

#4

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: Takagi et al.

Date: September 25, 2001

Serial No.: 09/682,145

Group Art Unit: 2152

Filed: July 26, 2001

Examiner:

For: System And Media For Simplifying Web Contents, And Method Thereof

Assistant Commissioner for Patents
Washington, D.C. 20231

SUBMISSION OF PRIORITY DOCUMENT

Sir:

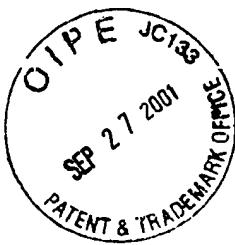
Enclosed herewith is a certified copy of Japanese Application No. 2000-227996 filed July 27, 2000, in support of applicant's claim to priority under 35 U.S.C. 119.

Respectfully submitted,



Derek S. Jennings
Reg. Patent Agent/Engineer
Reg. No.: 41,473
Tel. No.: (914) 945-2144

IBM CORPORATION
Intellectual Property Law Dept.
P. O. Box 218
Yorktown Heights, N. Y. 10598



日本国特許庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日

Date of Application:

2000年7月27日

出願番号

Application Number:

特願2000-227996

出願人

Applicant(s):

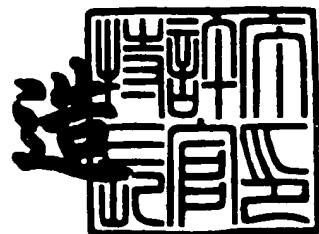
インターナショナル・ビジネス・マシーンズ・コーポレーション

CERTIFIED COPY OF
PRIORITY DOCUMENT

2001年1月12日

特許庁長官
Commissioner,
Patent Office

及川耕



【書類名】 特許願

【整理番号】 JP9000237

【提出日】 平成12年 7月27日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 19/00

【発明者】

【住所又は居所】 神奈川県大和市下鶴間1623番地14 日本アイ・ビ
ー・エム株式会社 東京基礎研究所内

【氏名】 ▲高▼木 啓伸

【発明者】

【住所又は居所】 神奈川県大和市下鶴間1623番地14 日本アイ・ビ
ー・エム株式会社 東京基礎研究所内

【氏名】 浅川 智恵子

【特許出願人】

【識別番号】 390009531

【氏名又は名称】 インターナショナル・ビジネス・マシーンズ・コーポレ
ーション

【代理人】

【識別番号】 100086243

【弁理士】

【氏名又は名称】 坂口 博

【復代理人】

【識別番号】 100112520

【弁理士】

【氏名又は名称】 林 茂則

【電話番号】 046-277-0540

【選任した代理人】

【識別番号】 100091568

【弁理士】

【氏名又は名称】 市位 嘉宏

【選任した代理人】

【識別番号】 100106699

【弁理士】

【氏名又は名称】 渡部 弘道

【選任した復代理人】

【識別番号】 100110607

【弁理士】

【氏名又は名称】 間山 進也

【選任した復代理人】

【識別番号】 100098121

【弁理士】

【氏名又は名称】 間山 世津子

【手数料の表示】

【予納台帳番号】 091156

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9706050

【包括委任状番号】 9704733

【包括委任状番号】 0004480

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 ウェブコンテンツを簡略化するための方法、システムおよび媒体

【特許請求の範囲】

【請求項 1】 簡略化対象の目的ページを取得するステップと、
前記目的ページに隣接する隣接ページを取得するステップと、
前記目的ページと前記隣接ページとで共通するオブジェクトを前記目的ページ
から削除する差分演算を施し、簡略化ページを生成するステップと、
を含むウェブコンテンツを簡略化するための方法。

【請求項 2】 前記隣接ページには、前記目的ページのURLまたは前記目的ページに含まれるリンクのURLとディレクトリが共通するURLのページ、または、

前記目的ページのURLまたは前記目的ページに含まれるリンクのURLと親ディレクトリが共通するURLのページ、または、

前記目的ページのURLを含むルートディレクトリ以下の各ディレクトリのトップページ、

を含む請求項1記載の方法。

【請求項 3】 前記隣接ページには、前記目的ページの過去のページにおいて前記過去ページに含まれていたリンクのページ、または、

前記目的ページまたは前記隣接ページの過去のページ、

を含む請求項1記載の方法。

【請求項 4】 前記隣接ページの取得後に前記隣接ページのURLの優先順位付けを行うステップをさらに有する請求項1記載の方法。

【請求項 5】 前記優先順位付けは、前記目的ページのURLと前記隣接ページのURLとの間のエディットディスタンス、または、前記目的ページと前記隣接ページとの間の共起回数または相互参照回数に基づくURL間の関連性、の何れかまたは両方に基づいて決定される請求項4記載の方法。

【請求項 6】 前記差分演算において、前記オブジェクトが共通するか否かの判断にはDPマッチングを用いる請求項1記載の方法。

【請求項7】 前記差分演算において、前記目的ページに含まれるオブジェクトの重要度を算出するステップを有し、

前記重要度が所定の閾値を超えている場合には前記オブジェクトが前記隣接ページのオブジェクトに共通する場合であっても前記オブジェクトを削除しない請求項1記載の方法。

【請求項8】 前記重要度は、重み付けされた特徴値の和で表し、

前記特徴値は、前記オブジェクトの文字サイズ、フォントその他の文字属性に割当てた数値、前記オブジェクトがバナーであることを識別判断するための数値、前記オブジェクトの画面中央位置からの変位値、前記オブジェクトに含まれるキーワード数、前記オブジェクトが追加または更新されたものであるかを示す情報に割当てた数値、前記オブジェクトの更新文字比率、前記オブジェクトが1文字であるかを示す情報に割当てた数値、前記オブジェクトのタグ種別に割当てた数値、の何れかである請求項7記載の方法。

【請求項9】 前記差分演算の後、前記簡略化ページに含まれる所定の閾値以下の重要度が低いオブジェクト、または、内容が空のテーブル要素またはリスト要素を削除するステップを有する請求項7記載の方法。

【請求項10】 前記差分演算の後、後処理ステップを有し、

前記後処理ステップでは、リストタイトルの修復、または、表頭および表側情報の修復、または、フォームのページ後部への移動、または、アノテーション情報の参照が行われる請求項1記載の方法。

【請求項11】 ユーザ端末からのリクエストを受け取るステップと、

前記リクエストに応答して前記各ステップを実行し、前記簡略化ページのうち、最も小さな情報量を有する簡略化ページを選択するステップと、

前記選択された最小の情報量を有する簡略化ページを前記ユーザ端末に送信するステップと、

を有する請求項1記載の方法。

【請求項12】 前記ユーザ端末は、音声ブラウザが稼動するコンピュータシステムまたは小画面の表示装置を有する情報端末である請求項11記載の方法

。

【請求項13】 前記ユーザ端末または前記ユーザ端末に接続されたコンピュータシステムは音声認識機能および音声合成機能を備え、音声による前記リクエストの入力ステップと、音声による前記簡略化ページの出力ステップと、を有する請求項11記載の方法。

【請求項14】 簡略化対象の目的ページを取得する手段と、前記目的ページの比較対象となる隣接ページのURLを生成する手段と、前記隣接ページを取得する手段と、前記目的ページおよび隣接ページに含まれる各々のオブジェクトを比較する手段と、前記オブジェクトの共通性を判断し、共通するオブジェクトを前記目的ページから削除して簡略化ページを生成する手段と、を含むウェブコンテンツを簡略化するためのシステム。

【請求項15】 前記隣接ページのURLを生成する手段には、前記目的ページのURLまたは前記目的ページに含まれるリンクのURLとディレクトリが共通するURLを生成する手段、または、前記目的ページのURLまたは前記目的ページに含まれるリンクのURLと親ディレクトリが共通するURLを生成する手段、または、前記目的ページのURLを含むルートディレクトリ以下の各ディレクトリを生成する手段、を含む請求項14記載のシステム。

【請求項16】 前記隣接ページのURLを生成する手段には、前記目的ページの過去のページにおいて前記過去ページに含まれていたリンクのURLを生成する手段、前記目的ページまたは前記隣接ページの過去のページのURLを生成する手段、を含む請求項14記載のシステム。

【請求項17】 前記隣接ページのURLの優先順位付けを行う手段をさらに有し、

前記優先順位付けは、前記目的ページのURLと前記隣接ページのURLとの間のエディットディスタンス、または、前記目的ページと前記隣接ページとの間の共起回数または相互参照回数に基づくURL間の関連性、の何れかまたは両方に基づいて決定される請求項14記載のシステム。

【請求項18】 前記オブジェクトの共通性の判断手段においてDPマッチングを計算する手段を含む請求項14記載のシステム。

【請求項19】 前記目的ページに含まれるオブジェクトの重要度を算出する手段を有し、

前記重要度が所定の第1閾値以上である場合には前記オブジェクトが前記隣接ページのオブジェクトに共通する場合であっても前記オブジェクトを削除しない手段と、

前記重要度が所定の第2閾値以下または前記オブジェクトの内容が空のテープル要素またはリスト要素である場合には前記オブジェクトを削除する手段と、

を有する請求項14記載のシステム。

【請求項20】 後処理手段をさらに有し、

前記後処理手段では、リストタイトルの修復、または、表頭および表側情報の修復、またはフォームのページ後部への移動、または、アノテーション情報の参照が行われる請求項14記載のシステム。

【請求項21】 ユーザ端末からのリクエストを受け取る手段と、

前記簡略化ページのうち、最も小さな情報量を有する簡略化ページを選択する手段と、

前記選択された最小の情報量を有する簡略化ページを前記ユーザ端末に送信する手段と、

をさらに有する請求項14記載のシステム。

【請求項22】 前記ユーザ端末は、音声ブラウザが稼動するコンピュータシステムまたは小画面の表示装置を有する情報端末である請求項21記載のシステム。

【請求項23】 前記ユーザ端末または前記ユーザ端末に接続されたコンピュータシステムは音声認識機能および音声合成機能を備え、

音声による前記リクエストの入力手段と、
音声による前記簡略化ページの出力手段と、
を有する請求項21記載のシステム。

【請求項24】 コンピュータ読み取り可能な記録媒体であって、コンピュータに、

簡略化対象の目的ページを取得する機能と、
前記目的ページに隣接する隣接ページを取得する機能と、
前記目的ページと前記隣接ページとで共通するオブジェクトを前記目的ページ
から削除する差分演算を施す機能と、
簡略化ページを生成する機能と、
を実現させるためのプログラムを記録した媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、ウェブコンテンツの簡略化方法およびシステムに関する。特に履歴情報を持たないウェブページや日々URL(Uniform Resource Locator)が変化するウェブページであっても、その内容をオンザフライで簡略化できる技術に関する。

【0002】

【従来の技術】

近年のネットワーク技術の進展と情報機器の機能の向上、低価格化とを反映して、インターネットの利用が盛んになっている。企業、個人を問わず低成本で、また国境を意識することなく詳細な情報発信を行えるため、情報発信源のウェブページは日々爆発的に増大している。また、ウェブページ管理者の管理の下に、日々膨大な情報がアップデートされている。このような背景の下にインターネットおよびこれを用いたウェブページは従来のブロードキャスト、マスメディアに代わる、あるいはこれを補う重要な情報収集媒体になりつつある。

【0003】

ところでウェブページの役割は多様化しつつある。たとえば単なる情報発信に

止まらず、ウェブページを介した商取引（電子商取引）、ウェブページを用いた共同作業（コラボレーション）等が行われつつある。このような多様化された機能を実現するため、より利便性の高いウェブページが提供される。また、目的の情報により迅速にアクセスするために検索画面等のユーザの操作性を向上する機能がウェブページに組み込まれる。たとえば、サイト内で共通に使われるリンクリストや、イメージマップ、フォーム等である。これらはどのページにも含まれており、一般ユーザにとって利便性の高い機能を提供する。

【0004】

しかし、これら一般のウェブページは、デスクトップ型のコンピュータ画面を前提として設計されている。つまりデスクトップ型コンピュータ画面のサイズを考慮してレイアウトされている。このため、PDA (Personal Digital Assistants) や携帯電話などの画面の小さなデバイス（以下、小画面デバイスと称する）や、ウェブページ読み上げソフトウェア（以下、音声ブラウザと称する）の場合には、素早く必要な情報にたどりつくことができないという問題がある。つまり、一般的なウェブページの場合、ページの上部にフォームやイメージマップがレイアウトされるため、小画面デバイスではこれらフォーム等の表示を何度も繰り返さなければ必要な情報に到達できない。また、音声ブラウザの場合にはこれらフォーム等の読み上げが行われた後に必要な情報の読み上げが行われる。一般に小画面デバイスの場合にはデスクトップ型コンピュータのようなビジュアルな多機能性は必要なく、また音声ブラウザの場合には操作性を向上させるビジュアルな機能は必要ない。これらビジュアル機能は小画面デバイスあるいは音声ブラウザには逆に邪魔になる。

【0005】

そこで、ウェブページの一部を省略する簡略化 (simplification) の手法が試みられる。たとえば、「<http://www9.org/w9cdrom/169/169.html>」に記載されているような”Dharma Transcoding”技術、あるいは、「<http://www.diffweb.com/>」に記載されているような”DiffWeb”（差分）技術がある。

【0006】

”Dharma Transcoding”技術は、すでに存在するウェブページを元のレイアウト

に近い状態でいくつかのページに分割し、小画面デバイスに表示しやすいページを作る技術である。この手法はページの構造や各部位の重要度などを詳細に記述した外部アノテーション情報を必要とする。

【0007】

"DiffWeb"技術は、あらかじめ登録し、保存されたウェブページと現在のウェブページの差分を計算、提示する技術である。ユーザ毎にページのリストを登録することができ、これらのページの差分を計算することができる。この差分技術では、ページの登録、保存、差分計算などすべての処理がユーザからの指示で行われる。同様の差分手法として、「<http://www-db.stanford.edu/c3/c3.html>」に記載された"HTML Diff"、【<http://mindit.netmind.com/mindit.shtml>】に記載された"MindIt"などもある。

【0008】

【発明が解決しようとする課題】

しかし、"Dharma Transcoding"技術では前記の通りアノテーション情報を必要とする。アノテーション情報の付与にはボランティア等の介在が必要であり、完全に自動化することは困難である。

【0009】

"DiffWeb"技術では、ページの登録、保存、差分計算が前記の通りユーザからの指示で処理される。このため、進行中(On The Fly)の処理として差分計算をすることができない。また、プルダウンメニューなどでは、中身の文字列が削除され簡略化後のフォームが正常に動作しない恐れがある。

【0010】

さらに、従来技術ではあらかじめ保存しておいた比較対象ページとの差分を計算することにより、シンプリフィケーションが実現されている。したがって、以下ののような問題がある。

【0011】

第1に比較対象ページがあらかじめ保存されていない場合には簡略化できないという問題がある。つまり比較ページを記録しているページのみが簡略化対象となり、初出ページの簡略化を実行できない問題がある。

【0012】

第2に、比較対象のページが保存されていた場合でも、URLが日々変化するページでは、簡略化できないという問題がある。たとえば朝日新聞(www.asahi.com)の記事ページは次のようにURLの中に日付が含まれている。

「<http://www.asahi.com/0530/news/business30010.html>」

この場合、同一URLの過去のページは存在せず、ページ簡略化を行うことができない。

【0013】

第3に必要な情報までも削除されてしまう問題がある。例えば、リンクリストのタイトルや、フォームといった重要な情報が削除されてしまう問題である。逆に不必要的微妙な文字列の変化などが保存されてしまう問題がある。

【0014】

本発明の目的は、小画面デバイスや音声ブラウザを用いたウェブページの表示あるいは出力の際に、必要な情報に迅速にアクセスするためのウェブページ簡略化の手法を提供することにある。

【0015】

また、本発明の目的は、ウェブページの簡略化を同一URLの過去のページが存在しない場合においても実行できる手法を提供することにある。

【0016】

また、本発明の目的は、ウェブページの簡略化をオンザフライで実行できる手法を提供することにある。

【0017】

また、本発明の目的は、ウェブページの簡略化の際に重要な情報を欠落することなく、不必要的情報を精度良く簡略化する手法を提供することにある。

【0018】

【課題を解決するための手段】

本願の発明の概略を説明すれば、以下の通りである。本発明は、差分計算に基づいて音声ブラウザや小画面デバイスで読みやすいページにオンザフライで変換する手法を提供する。本発明のページ簡略化で用いる差分計算では、簡略化対象

のページと比較する比較ページとして、同一URLの過去ページのみならず、隣接ページを用いる。比較により更新された情報のみを取り出し、サイト内の各ページで共通に使用されているテンプレート情報を取り除く。これにより簡略化対象ページの主なコンテンツのみを取り出す。

【0019】

本発明により、比較のための近隣ページを自動的に取得でき、過去ページの蓄積情報がなくともオンザフライでウェブページの簡略化が実行できる。これにより小画面デバイスあるいは音声ブラウザを用いた場合の必要な情報への迅速なアクセスが可能になる。

【0020】

具体的には本発明の方法は、簡略化対象の目的ページを取得するステップと、目的ページに隣接する隣接ページを取得するステップと、目的ページと隣接ページとで共通するオブジェクトを目的ページから削除する差分演算を施し、簡略化ページを生成するステップと、含む。

【0021】

前記隣接ページには、目的ページのURLまたは目的ページに含まれるリンクのURLとディレクトリまたは親ディレクトリが共通するURLのページを含み、目的ページのルートディレクトリ以下の各ディレクトリのトップページ、または、目的ページの過去のページ、過去ページに含まれていたリンクのページ、さらに隣接ページの過去のページを含む。

【0022】

前記隣接ページの取得後に前記隣接ページのURLの優先順位付けを行うことができ、優先順位付けは、目的ページのURLと隣接ページのURLとの間のエディットディスタンス、または、目的ページと隣接ページとの間の共起回数または相互参照回数に基づくURL間の関連性に基づいて決定できる。

【0023】

前記差分演算において、オブジェクトが共通するか否かの判断にはD Pマッチングを用いることができ、また、目的ページに含まれるオブジェクトの重要度を算出できる。重要度が所定の閾値を超えている場合にはオブジェクトが隣接ペー

ジのオブジェクトに共通する場合であってもオブジェクトを削除しない。逆に重要度の低いオブジェクトは削除できる。

【0024】

前記重要度は、重み付けされた特徴値の和で表し、特徴値には、オブジェクトの文字サイズ、フォントその他の文字属性に割当てた数値、オブジェクトがバーであることを識別判断するための数値、オブジェクトの画面中央位置からの変位値、オブジェクトに含まれるキーワード数、オブジェクトが追加または更新されたものであるかを示す情報に割当てた数値、オブジェクトの更新文字比率、オブジェクトが1文字であるかを示す情報に割当てた数値、オブジェクトのタグ種別に割当てた数値、等を採用できる。

【0025】

また、差分演算の後、後処理を行うことができ、後処理では、リストタイトルの修復、または、表頭および表側情報の修復、または、フォームのページ後部への移動、または、アノテーション情報の参照が行える。

【0026】

また、ユーザ端末からのリクエストを受け取り、リクエストに応答して前記各ステップを実行し、簡略化ページのうち、最も小さな情報量を有する簡略化ページを選択してユーザ端末に送信することができる。ユーザ端末は、音声ブラウザが稼動するコンピュータシステムまたは小画面の表示装置を有する情報端末とすることができる。あるいはユーザ端末またはユーザ端末に接続されたコンピュータシステムに音声認識機能および音声合成機能を備え、音声によるリクエストの入力、音声による簡略化ページの出力を行うことができる。

【0027】

【発明の実施の形態】

以下、本発明の実施の形態を図面に基づいて詳細に説明する。ただし、本発明は多くの異なる態様で実施することが可能であり、本実施の形態の記載内容に限定して解釈すべきではない。なお、実施の形態の全体を通して同じ要素には同じ番号を付するものとする。

【0028】

以下の実施の形態では、主に方法またはシステムについて説明するが、当業者であれば明らかなどおり、本発明は方法、システムの他、コンピュータで使用可能なプログラムが記録された媒体としても実施できる。したがって、本発明は、ハードウェアとしての実施形態、ソフトウェアとしての実施形態またはソフトウェアとハードウェアとの組合せの実施形態をとることができる。プログラムが記録された媒体としては、ハードディスク、CD-ROM、光記憶装置または磁気記憶装置を含む任意のコンピュータ可読媒体を例示できる。

【0029】

また以下の実施の形態では、一般的なコンピュータシステムを用いることができる。実施の形態で用いるコンピュータシステムには、中央演算処理装置（CPU）、主記憶装置（メインメモリ：RAM）、不揮発性記憶装置（ROM）等を有し、バスで相互に接続される。バスには、その他コプロセッサ、画像アクセラレータ、キャッシュメモリ、入出力制御装置（I/O）等が接続されてもよい。バスには、適当なインターフェイスを介して外部記憶装置、データ入力デバイス、表示デバイス、通信制御装置等が接続される。その他、一般的にコンピュータシステムに備えられるハードウェア資源を備えることが可能なことは言うまでもない。外部記憶装置は代表的にはハードディスク装置が例示できるが、これに限られず、光磁気記憶装置、光記憶装置、フラッシュメモリ等半導体記憶装置も含まれる。データ入力デバイスには、キーボード等の入力装置、マウス等ポインティングデバイスを備えることができる。データ入力デバイスにはスキャナ等の画像読み取り装置、音声入力装置も含む。表示装置としては、CRT、液晶表示装置、プラズマ表示装置が例示できる。また、コンピュータシステムには、パーソナルコンピュータ、ワークステーション、メインフレームコンピュータ等各種のコンピュータが含まれる。

【0030】

（実施の形態1）

図1は、本発明の実施の形態1のシステムを示した構成図である。本実施の形態のシステムは、ユーザ端末1、プロキシサーバ2、目的のウェブページが存在するウェブサーバ3を含む。

【0031】

ユーザ端末1は音声ブラウザが稼動するコンピュータシステムまたは小画面の表示デバイスを有する小画面デバイスである。音声ブラウザはHTML (Hypertext Markup Language) 文書またはXML (eXtensible Markup Language) 文書等HTTP (HyperText Transfer Protocol) を用いて取得される情報(文書)を音声によりブラウジングする機能を有するソフトウェアである。また、小画面デバイスは、たとえばiモード携帯電話やPDA等デスクトップ型コンピュータの表示装置よりも格段に小さい表示画面を有する情報端末である。これら音声ブラウザを用いて取得した文書を読み上げる際、あるいは小画面デバイスに表示する際には、デスクトップ型コンピュータ用にレイアウトされたウェブコンテンツをそのまま適用したのでは読み上げあるいは表示の障害になる。本実施の形態のシステムはこのようなユーザ端末1においてもスムーズに取得情報の読み上げあるいは表示を行うことを目的としている。特に音声ブラウザのユーザとなるであろう視覚障害者にとっては、視覚障害をもたない者と同様な情報取得手段を得ることになる。本発明は情報アクセスへの平等な機会を確保できる手段を提供でき、社会的な意義も大きい。

【0032】

プロキシサーバ2は、一般的なプロキシサーバの機能に加えて、ウェブページを簡略化する差分処理4の機能を有する。差分処理4の機能については後に説明する。差分処理4においてはキャッシュデータベース5を参照する。キャッシュデータベース5には、過去においてプロキシサーバ2がアクセスしたウェブページが格納される。なお、本実施の形態ではプロキシサーバ2を用いて簡略化されたウェブページを表示または音声出力する例を説明するが、簡略化機能(差分処理4の機能)を各クライアント(ユーザ端末1)あるいはウェブサーバ3に有しても良い。また、キャッシュデータベース5はプロキシサーバ2のシステム内に存在するように図示しているが、キャッシュデータベース5は必ずしもプロキシサーバ2内に存在する必要はない。たとえばURL、IPアドレス等でその所在が特定でき、サーバの稼動状態においてアクセスできる限り他のシステムに存在しても良い。他のシステムにはLAN又はWAN接続される他のシステム、イン

ターネットを介して接続される他のシステムが含まれる。また、キャッシュデータベース5は概念的なものであり、物理的には異なる場所（アドレス）に存在する記憶装置によって分散的に処理されても良い。以後説明する他のデータベースについて同様である。

【0033】

ウェブサーバ3には目的のウェブページ6を有する。一般にウェブサーバ3はインターネットに接続されていることが想定されるが、必ずしもインターネットに接続されている必要はなく、HTTPリクエストに対するレスポンスを行うサーバであれば良い。たとえば企業等団体内のイントラネット、エクストラネット等各種ネット上のサーバでも良い。

【0034】

また、ウェブサーバ3には目的ページ6に隣接する隣接ページ7が存在する。隣接ページ7については後に詳述する。通常隣接ページ7には目的ページ6と共通なフォーム、リンクリスト、イメージマップ等のオブジェクトを含み、これら共通のオブジェクトを目的ページから削除することによりウェブページの簡略化を行う。なお、本実施の形態において隣接ページ7とともに比較対照ページの候補としてキャッシュデータベース5に記録されている過去ページも用いるが、過去ページについては本発明に必須の構成要件ではない。本発明では目的ページ6の過去ページがなくとも現在の隣接ページ7を用いて簡略化を実行できる。

【0035】

本実施の形態のシステムでは、ユーザはユーザ端末1のブラウザのプロキシサーバの設定においてプロキシサーバ2のアドレスを指定する。そしてウェブサーバ3にアクセスすることにより簡略化されたウェブページ8を表示あるいは出力できる。

【0036】

本システムの処理の概要は以下の通りである。ユーザ端末1は、ウェブサーバ3のアドレスを指定してHTTPリクエスト9を出す。ユーザ端末1からのリクエスト9を受けてプロキシサーバ2がウェブサーバ3にHTTPリクエスト10を出す。プロキシサーバ2は目的ページ6、隣接ページ7およびキャッシュデータ

タベース5内の必要に応じた過去ページの参照により差分処理4を行い、簡略化ページ8をリクエスト9の応答として返送する。ユーザ端末1は簡略化ページ8を表示し、あるいは音声出力を行う。このような簡略化方法により、目的ページ6の共通のオブジェクトを削除し、ユーザは迅速に必要な情報を得ることができる。簡略化されたページを読み上げるためユーザはよりスムーズに、ストレスを感じることなく必要な（重要な）情報を聞くことができる。特に視覚障害者には視覚的にウェブページを把握することができないので、本発明を用いたウェブページの簡略化により、スムーズに情報を取得できる。以下、差分処理の詳細を説明する。

【0037】

図2は差分処理4の詳細を示したブロック図である。本実施の形態の差分処理モジュールには、隣接URL列挙11、ディレクトリ列挙12、URLキャッシュ13、更新前目的ページ/隣接ページ取得14の各モジュールと、ウェブページを取得しDOM (Document Object Model) によるオブジェクトを出力するフェッチモジュール、URL優先順位演算モジュール16、差分演算モジュール17、最小差分選択モジュール18を含む。本実施の形態の差分処理は、目的のページに対して、差分計算の比較対象として適切なページを自動的に検出し、差分を算出する。本実施の形態では、隣接ページという概念を導入することで、過去のページが保存されていない場合、あるいは日々URLが変化する場合でも差分の抽出を可能とする。同時に、サイト内で共通に使用されるテンプレート（ヘッダー、フォーム、リンクリストなど）を任意のページに対して取り除くことができるようになる。隣接ページとは、サーバ上で同じディレクトリに存在するページ、親ディレクトリが同じページ、ディレクトリのトップページ、サイトのトップページなど目的ページに含まれるオブジェクトと共通なオブジェクトを含む可能性のあるページをいう。本実施の形態では、これらの隣接ページとの差分も計算し、その中でもっとも文字列の短くなる差分を選択することで、オンザフライ（進行中）でのウェブページ簡略化を実現する。

【0038】

隣接URL列挙モジュール

図3は、隣接URL列挙モジュール11の詳細を示したブロック図である。隣接URL列挙モジュール11は、サイト内の近い場所にあるページ（隣接ページ7）のリストを、目的ページ6を解析して得るモジュールである。これらのページ（隣接ページ7）には目的ページ6と同じリンクリスト、イメージマップを持っていることが多いため、これら隣接ページの共通オブジェクトを削除して目的ページ6の簡略化を実施できる。

【0039】

隣接URL列挙モジュール11には、リンク列挙モジュール19、URLリスト20、ディレクトリ共通URL選択モジュール21、親ディレクトリ共通URL選択モジュール22、リストマージモジュール23を含む。

【0040】

まず、プロキシサーバ2からのリクエストによって、目的ページ6からコンテンツをフェッチする（図2のフェッチモジュール15）。フェッチによって取得された目的ページのオブジェクト（DOM）から、目的ページ6内に含まれるURLを列挙する（リンク列挙モジュール19）。URLの列挙には、たとえば[タグのhrefアトリビュートを参照する。そしてリンク列挙モジュール19の出力としてURLリスト20を得る。URLリスト20の中から、目的ページ6のURLと同一のディレクトリをさしているURLを選択し、新たなリストを作成する（ディレクトリ共通URL選択モジュール21）。さらに、親ディレクトリが共通なURLを選択してURLのリストを作成する（親ディレクトリ共通URL選択モジュール22）。最後にこれらのリストをマージモジュール23によりマージして、隣接URLリスト24を得る。後に説明するようにそれぞれのURLに対してフェッチモジュール15を起動して差分演算に用いる。](#)

【0041】

21、22のいずれのモジュールによるリストがより有効かは各サイトの構造により異なる。例えば、"www.asahi.com"では、ディレクトリ共通URL選択モジュール21が有効であり、"www.cnn.com"では親ディレクトリ共通URL選択モジュール22が有効である。このようにサイトの構造によって有効な方法が変化するため、両方のリストをリストアップして、より有効なページ簡略化の比較

ページを得ることができる。

【0042】

以下にURLの選択例を示す。

(1) ディレクトリ共通URL選択例：

目的ページのURL：

<http://www.asahi.com/0606/news/national06015.html>

リストアップされるURL（一部）：

<http://www.asahi.com/0606/news/national06012.html>

<http://www.asahi.com/0606/news/national06013.html>

<http://www.asahi.com/0606/news/national06014.html>

(2) 親ディレクトリ共通URL選択例：

目的ページのURL：

http://www.cnn.com/2000/US/06/05/sea_based_defense/index.html

選択されるURL（一部）：

http://www.cnn.com/2000/US/06/05/dday_remembrance/index.html

http://www.cnn.com/2000/US/06/05/helicopter_escape_03/index.html

http://www.cnn.com/2000/US/06/05/curbing_terrorism_02/index.html

【0043】

ディレクトリ列挙モジュール

ディレクトリ列挙モジュール12は、目的ページ6のURLから各ディレクトリのトップページのリストを作成する。大規模なサイトにおいて、いくつかのトップページに分かれることがあり、目的ページと共通のリンクリストを含む場合の比較対照ページの取得に有効である。以下に具体例を例示する。

目的ページのURL：

http://www.cnn.com/2000/US/06/05/helicopter_escape_03/index.html

選択されるURL：

<http://www.cnn.com/2000/US/06/05/>

<http://www.cnn.com/2000/US/06/>

<http://www.cnn.com/2000/US/>

<http://www.cnn.com/2000/>

<http://www.cnn.com/>

【0044】

URLキャッシュモジュール

図4はURLキャッシュモジュール13の一例を示した構成図である。URLキャッシュモジュール13は、URLキャッシュデータベース25、検索手段26を含む。URLキャッシュデータベース25は図1のキャッシュデータベース5の一部として構成されても良い。URLキャッシュデータベース25には、過去にプロキシサーバ2のユーザがアクセスしたページのURLとそのページ内に含まれているURLのリストが記録されている。検索手段26は目的ページ6のURLに応じてキャッシュ内のURLを検索し、目的ページ6と同一サイトのURLリスト27をリストアップする。

URLキャッシュモジュール13によりURLキャッシュから目的ページのURLに近いURLを取得することができ、適切なページの候補を増やすことが可能になる。

【0045】

URL優先順位演算モジュール

URL優先順位演算モジュール16は、前記した隣接URL列挙モジュール11、ディレクトリ列挙モジュール12、URLキャッシュモジュール13によってリストアップされたURLの類似度を算出し、URLの優先順位付けを行うモジュールである。優先順位の順に後に説明するページのフェッチが行われる。URL優先順位演算モジュール16は、より確実にレイアウトの似たページをリストアップするために、URLのエディットディスタンス(Edit distance)、共起回数、相互参照回数を総合して決定する。

図5は、URL優先順位演算モジュール16の一例を示したブロック図である。URL優先順位演算モジュール16には、同一ページ除外モジュール28、URLエディットディスタンス算出モジュール29、URL関連性算出モジュール30、ソートモジュール31、リダイレクションURLテーブル32、URL共起回数テーブル33、URL相互参照テーブル34を含む。

【0046】

次に、URL優先順位演算モジュール16の動作を説明する。まず、隣接URL列挙モジュール11、ディレクトリ列挙モジュール12、URLキャッシュモジュール13でリストアップされたURLを統合し、一つのURLリストを作る。次にリダイレクションURLテーブル32を用いて、同一ページを指し示しているURLを同一ページ除外モジュール28で除外する。リダイレクションURLテーブル32は、後に説明するように各フェッチモジュール15によってメンテナンスされる。

【0047】

次に、目的ページ6のURLとリスト中の各URLとの類似度を算出する。類似度はURLエディットディスタンス算出モジュール29を用いてエディットディスタンス (edit distance) を算出することにより算出できる。例えば、DPマッチング (DP Matching) を用いて最長共通文字列 (Longest common string) を算出し、そこから編集操作回数を算出してエディットディスタンスを求める。エディットディスタンスが短いほど類似度は高い。

【0048】

次に、URL関連性算出モジュール30を用いてURLの関連性を算出する。URL関連性算出モジュール30は、エディットディスタンスの値と、URLの相互参照回数、URLの共起回数とを総合して、目的ページのURLとの関連性を数値化する。URLの共起回数および相互参照回数は、各々URL共起回数テーブル33、URL相互参照回数テーブル34に記録され、フェッチモジュール15内のHTMLパーザ (後に説明する) において常に算出、更新される。URL関連性算出モジュール30は、URLの相互参照回数、共起回数、エディットディスタンスを重みつき加算し、各URLの関連性を算出する。

最後に、ソートモジュール31において降順でリストを並び替え、優先順位を決定する。決定された優先順位は優先度順URLリスト35として出力される。

【0049】

更新前目的ページ/隣接ページ取得モジュール

更新前目的ページ/隣接ページ取得モジュール14は、目的ページ6またはそ

れに隣接するページ7の更新前ページがキャッシュデータベース5に存在する場合にはこれらをも差分(比較)対象として選択するためのモジュールである。

図6は、更新前目的ページ/隣接ページ取得モジュールの一例を示した構成図である。更新前目的ページ/隣接ページ取得モジュール14には、検索キー作成モジュール36、検索モジュール37、HTMLパーザ39を含み、前記キャッシュデータベース5が参照される。

【0050】

URL優先順位演算モジュール1-6によって、類似度順にソートされたURLリスト35の過去のページがキャッシュデータベース5内に存在するかを検索モジュール37を用いて検索する。存在していた場合、キャッシュデータベース5から検索条件を満足するウェブページのリスト38を抽出し、HTMLパーザ39を用いてHTML文書からウェブページリスト40(DOMツリーからなる)を生成する。キャッシュデータベース5にはURLリストだけでなく、ウェブページのコンテンツも記録されるので、更新前目的ページ/隣接ページリスト41はDOMツリーの状態で得られる。得られた更新前目的ページ/隣接ページリスト41は目的ページ6の差分対象として選択する。

【0051】

キャッシュデータベース5に対する検索キーは検索キー作成モジュール36で生成される。検索キーには、目的ページ6のURLに加えて、前述のディレクトリ列挙モジュール12によって生成されるURL列を用いる。

また、目的ページ6が検索エンジン(ホームページ検索を主な機能とするウェブページ)の検索結果など「Query文字列」を含む場合、他のキーワードによる検索結果(URL)も検索キーに加えることができる。

【0052】

フェッチモジュール

図7は、フェッチモジュールの一例を示す構成図である。フェッチモジュール15には、ダウンロードモジュール42およびHTMLパーザ43を含む。フェッチすべきURL44の入力を受けて、ダウンロードモジュール42は入力されたURLのウェブサーバ3にHTTPリクエストを発行する。ウェブサーバ3は

リクエストを受けてHTMLファイル45を返送する。ダウンロードモジュール42は受け取ったHTMLファイルのURLをリダイレクションURLテーブル32に記録する。一方、HTMLファイル45はHTMLパーザ43に送付され、DOMツリー46に変換される。HTMLパーザ43は、URL毎の共起回数、相互参照回数を算出し、URL共起回数テーブル33およびURL相互参照テーブル34に記録する。これらリダイレクションURLテーブル32、URL共起回数テーブル33およびURL相互参照テーブル34はフェッチモジュール15によってメンテナスされ、前記URL優先順位演算モジュール16で利用される。

【0053】

差分演算モジュール

差分演算モジュール17は、目的ページ6のDOMツリーと前記手法で選択された比較対象ページのDOMツリーとから差分のDOMツリーを生成する。

図8は差分演算モジュールの一例を示したブロック図である。差分演算モジュール17には、線形化モジュール47、DPマッチングモジュール50、重要度算出モジュール52、共通ノード削除モジュール53を含む。URL優先順位演算モジュール16によって並び替えられた各URLは、前記したフェッチモジュール15によってDOMツリーに変換される。一方、更新前目的ページ/隣接ページ取得モジュール14によって取得されたリスト41はDOMツリー状態で得られるためフェッチの必要はない。

【0054】

これら隣接URL列挙(11)、ディレクトリ列挙(12)、URLキャッシュ(13)、更新前目的ページ/隣接ページ取得(14)の各モジュールで選択されたリストに対応するページのDOMツリーを前記の通り生成し、これを差分演算の一方の入力として目的ページ6の比較対象とする。目的ページ6のDOMツリーは差分演算の他の入力となる。

【0055】

本実施の形態の差分過程で重要なノードが削除されるのを防ぐために、あらかじめノードの重要度を重要度算出モジュール52によって算出し、ある閾値を越

えた重要度を持つノードは共通ノードであっても削除しない。また、次のクリーンアップモジュール54で重要度の低いノードが取り除かることでそのページに独自でかつ、重要度の高い情報のみを取得できる。

【0056】

以下、DPマッチングを用いた手法を図8に従って説明する。まず、線形化モジュール47が、目的ページと比較対象ページそれぞれのDOMツリーからノードリスト48、49を作成する。線形化モジュール47はDOMツリーを巡回して、テキストノード、イメージ(画像)ノードを選択する。このとき、ページ上のフォームが削除されないように、フォームノードの下のノードを選択せず、フォームの処理を後に他の実施の形態として説明する「後処理」モジュールにおいて行うようにすることもできる。また、クライアントサイドスクリプト(JavaScript, VBScript)を保存するためにコメントノードも選択しないようにすることができる。

【0057】

次に、2つのノードリストに対してDPマッチング50を行い、共通に含まれるノードのリストを算出する。DPマッチングは2つの記号列から、共通するもっとも長い記号列(Length Common String)を算出するためのアルゴリズムである。例えば、"abcdefghijklm"と"bcdegh"であれば"bcdgh"が outputされる。本実施の形態ではこのアルゴリズムをノードのリストに対して適応することで共通に含まれるノードのリスト(いわばLength Common "Node" String)を得る。

【0058】

重要度算出モジュール

差分演算において、重要ノード(タイトルを示す文字列など)が削除されるのを防ぐため、あらかじめ目的のページの各テキストノードおよびイメージノードに対し、重み付けを行う。共通ノード削除モジュール53では、重みが閾値を越えていた場合、削除しない。

ノードの重要度を算出する手法を以下に例示する。なお以下に例示する以外にその他の重要度算出手法を適用できることは勿論である。ここではいくつかの特徴値の重みつき総和により重要度を決定する手法を説明する。各ノードの重要度

s は次の式により算出される。

$$s = \sum W_i * P_i$$

ただし、 P_i = 各特徴値

W_i = 各特徴値に対する重み付け

【0059】

以下に特徴値の例を述べる。

[文字サイズ] レンダリングされたときの文字サイズ (size) とデフォルトフォントサイズ (default size) の差を特徴値 P_i とする。

$$P_i = size - \text{default size}$$

大きな文字サイズを有するノードほど重要度が高いという経験則に基づく。文字属性を特徴値 P_i に加味することもできる。その場合、各属性値に応じて P_i に加算される。たとえばボールド、イタリック等のフォント、赤等の色が指定されている場合、下線、二重下線等の属性が指定されている場合には一般に重要度が高いと判断でき、このような属性に応じて P_i に加算する。

【0060】

[テンプレートによるバナーの除去] バナーである可能性が高いイメージリンクの場合には重要度を下げる。バナーテンプレートは、画像サイズ、リンク先文字列 (/doubleclick/, /ads/ など)、直後のリンク文字列 (Click Here) などを判断基準とすることができます。テンプレートからの距離を特徴値 P_i とすることができる。

【0061】

[ノードのポジション] レンダリングされたときにノードが表示されるポジションにより、重み付けを行う。図9に示すように中心部のものほど位置重要度を上げ、周辺のものほど重要度を下げる。図9ではウインドウ内の色の濃い領域ほど重要度が高いことを示している。特徴値 P_i は、各ノードの各ピクセル値の位置重要度の総和で算出できる。

【0062】

[キーワード検出による重要度上げ] 目的ページのキーワード解析を行い、キーワードを含むノードの重要度を上げることができる。システムが持っている

重要キーワードと、ページ内を解析して決定されたキーワードがノード内に含まれている個数をそのノードの特徴値とすることができます。

【0063】

[追加されたノードと更新されたノード] 追加されたノード（比較対象のページには含まれないノード）の重要度を上げるために、追加されたノードに対して特徴値を1、それ以外は0とすることができる。Wiは正の値とする。

【0064】

[更新されたノードの場合更新文字列比率] 追加ではなく、更新されたノードの場合、ノード内の文字数に対する更新文字数の割合を特徴値とすることができる。Wiは正の値とする。

【0065】

[1文字の時は下げる] 一字のみのノードの重要度を下げるため、特徴値1を割りあてることができる。この時Wiは負の値とする。

【0066】

[タグ種別] ノードによっては、タグの種類によって明確に重要度が判別できるものがある。そのようなタグに特徴値を割り当てる。デフォルトは0とする。たとえば、フォームノードを保存するためにFORMノードに正の特徴値を割り当てることができる。

【0067】

共通ノード削除モジュール

共通ノード削除モジュール53は、共通ノードリスト51に含まれるノードを目的ページ6のDOMツリーから削除することで差分のDOMツリーを生成する。ただし、重要度算出モジュール52で重要度が高いと判断されたノードは削除しない。保存するノードは、一定の閾値によって決定される。閾値はシステムの持つデフォルト値以外に、ユーザによって指定される閾値も利用可能である。出力は差分DOMツリーとなる。

【0068】

クリーンアップモジュール

クリーンアップモジュール54は、差分処理の最後で、重要度の低いノードと

空白ノードを削除する。まず、重要度算出モジュール52においてきわめて低い重要度であると判断されたノードを削除する。削除されるノードはある閾値によって判別される。この閾値はシステムの持つデフォルト値以外に、ユーザによって指定される閾値も利用可能である。次に、空白のテーブルセル<TD>、リストアイテムなどを削除する。テーブルセルは、列または行のすべてが空白の場合にのみ削除を行う。

【0069】

最小差分選択モジュール

最小差分選択モジュール18は、各比較ページに対応する差分演算モジュール17の出力から最小サイズの差分を選択する。最も効率的に簡略化された差分ページをプロキシサーバ2の出力（簡略化ページ8）としてユーザ端末1に返送し、ユーザ端末1は簡略化ページ8をブラウジングする。

【0070】

本実施の形態のシステムおよび簡略化方法によれば、過去ページが存在しない場合であっても、比較対象のページを得ることができ、目的ページの簡略化を行うことができる。また、各種の隣接ページ（比較対象ページ）を網羅的に取得するため、より適切な精度の高い簡略化を行える。また、差分処理においてノードの重要度をチェックするため必要な情報を欠落させる確率が低くできる。また、クリーンアップモジュールを用いるため無用なノードを削除でき、より重要な情報のみを残して簡略化の精度を向上することができる。

【0071】

実際のウェブページに本実施の形態のシステムおよび簡略化方法を適用した例を示す。図10は通常のブラウザによりニュースページを表示した表示図である。つまり図10は本実施の形態の処理を施す前の表示に対応する。図11は、本実施の形態のシステムを用いて図10のページを表示した表示図である。ページ上部および左側に存在していたリンクリストが削除され、ページ固有の情報であるニュース本文が残っていることがわかる。音声ブラウザ等でブラウジングした時により速やかにニュース本文に到達できることがわかる。

【0072】

図12は検索画面の一例を通常のブラウザにより表示した例を示す表示図である。検索結果が画面中央部に表示されている。図13は、本実施の形態のシステムを用いて図12のページを表示した表示図である。前記図11の場合と同様にリンクリストが削除され、検索結果が残されている。なお、画面左に位置する検索機能をもつフォームはその位置にリンクだけが残され、フォーム自体はページの後部に移動される。これにより音声ブラウザによる読み上げが検索結果に素早く到達できる。フォームの移動については後に説明する。

【0073】

前記した本実施の形態のシステムおよび簡略化方法により、ページの文字数、リンク数、ページ内エレメント数ともに半分程度に削減できることが本発明者の検討により判明した。表1は、CNN、朝日新聞、SUNTIMESの任意のページにおいて、本実施の形態のシステムおよび方法を適用した結果を示す表である。ばらつきがあるものの、おおむねオリジナルページに対して40%から60%までに情報を削減できていることがわかる。

【0074】

【表1】

Site	Number of Characters		Number of Links		Number of Elements	
	original	transcoded	original	transcoded	original	transcoded
CNN	4294	2557	60%	167	75	45%
Suntimes	3446	2770	80%	59	17	29%
Asahi	1880	1020	54%	40	4	10%

また、表2は各種検索ページの開始から検索結果を表示するまでの情報量の比較を示した表である。情報が大幅に削減されており、音声ブラウザなどでよりすばやく検索結果に到達できるようになっている事がわかる。

【0075】

【表2】

Page	Original	Transcoded
Yahoo	14 links 1 image map	0 link
Lycos	15 links 1 form	7 links
Infoseek	16 links and 1 form	2 links

(実施の形態2)

図14は、本発明の実施の形態2のシステムを示した構成図である。本実施の形態のシステムは、実施の形態1のシステムを構成する各構成要件に加えてプロキシサーバ60に後処理モジュール61、DOMからHTMLへの変換モジュール62、ユーザプロファイル63を有する。本発明の実施の形態のシステムは、特に視覚障害者のユーザに適する。視覚障害者のユーザがアクセスすることによって自動的にページを読みやすく変換して出力するプロキシサーバ60を提示できる。このようなプロキシサーバ60は、小画面デバイスや、電話によるWEBアクセスのサーバへの応用も勿論可能である。

【0076】

図14に示すように、ユーザはユーザ端末1のブラウザのプロキシサーバとして本システムのプロキシサーバ60のアドレスを指定する。ユーザプロファイル63により、ユーザの要求に合わせた簡略化を行える。たとえば、ユーザがあるページに対して、より絞り込んだ情報のみを要求した場合、本システムは共通ノード削除モジュール53(図8)およびクリーンアップモジュール54(図8)における、閾値の値を上げることで情報をより重要度の高いものだけに絞り込んで提示することができる。ユーザによる閾値コントロールは、各ページの最下部に「情報量減らす」と「情報量増やす」という二つのリンクを追加し、これらのリンクが選択されたときに閾値を変更するという方法で実現できる。同様に他のパラメータをユーザがコントロールすることも考えられる。このようなカスタマイゼーション機能は、本発明に容易に統合することが可能である。

【0077】

また、差分DOMツリー55に対していくつかのヒューリスティクスを用いて

一部の情報をさらに修復することでさらに精度を上げることが可能である。後処理モジュール61では、このようなヒューリスティクスに基づいた処理を行うことができる。

【0078】

図15は後処理モジュール61の一例を示したブロック図である。ここでは、タグ構造の分析に基づいた差分結果（差分処理の出力）の自動修正の例を示す。後処理モジュール61には、リストタイトル修復モジュール64、表頭・表側修復モジュール65、フォーム移動モジュール66を有する。

【0079】

リストタイトル修復モジュール64では、リスト（HTMLのオーダードリストとアンオーダードリスト）のタイトルが削除されてしまっていた場合にこれを目的ページのDOMツリーを参照して修復する。図16はリストタイトル修復モジュール64による修復の例を示す表示図である。オリジナル（同図（a））の「同ジャンルの他のニュース」というリンクリストのタイトルを示す文字列が差分ページ（同図（b））では削除されている。このようなリンクリストのタイトルは各ページに共通に含まれる場合が多いため、このような現象が生じる可能性がある。リンクリストのタイトルはそのリストの意味を的確に表しており、残されるべきである。そこで、リストタイトル修復モジュール64では、次のような条件でタイトルを識別、修復する。

- 1) リスト内の項目が一つでも残っている。
- 2) リストの直前の文字列がヘッダー、ボールド、拡大フォントのいずれかでかつ、50文字以内である。

このような場合、直前の文字列をタイトルと判定し、同図（c）に示すように、修復する。

【0080】

表頭・表側修復モジュール65においては、同様にテーブル内のセルが残っていた場合にテーブルの先頭のセルを修復する。

【0081】

フォーム移動モジュール66では、図17に示すように、フォームをページの

下部に移動する。図17(a)はフォームの存在した位置に移動したフォームへのリンクを残した図であり、図17(b)はページ下部に移動されたフォームを示す図である。ページ上部にあるフォームは音声ブラウザを用いた場合大きな障害となる。一方、検索フォームのように重要なフォームはページに残すことがほしい。そこでフォーム移動モジュール66では、フォームをページ下部へ移動するとともに、そのフォームへのリンクをフォームが存在した場所に残すことにより、簡略化とフォームの保存を両立させることができる。

【008-2】

なお、本実施の形態では実施の形態1とは相違して差分ページ(DOM)をDOMからHTMLへの変換モジュール62によりHTMLに変換する。得られる差分ページはHTMLファイルである(67)。このような場合、ダイナミックHTMLの対応しないブラウザでの表示が可能になる。

【008-3】

(実施の形態3)

図18は、本発明の実施の形態3のシステムを示した構成図である。本実施の形態のシステムは、アノテーション情報による差分ページの修復、修正を可能にする例である。

【008-4】

従来技術の項で前記した通り、詳細なアノテーション情報に基づいて小画面デバイス用の画面出力を得る手法が提案・開発されている。本実施の形態のシステムでは、このようなアノテーション情報と組みあわせることで、より精度の高い出力を得ることができる。ここではアノテーション情報が後処理において用いられる例を示す。図18に示すように、本実施の形態のプロキシサーバ70には実施の形態2のプロキシサーバ60の構成(ユーザプロファイル63を除く)に加えてアノテーション用の後処理モジュール71を有する。また本実施の形態のシステムはアノテーションデータベース72を有する。

【008-5】

目的ページ6に対して、ボランティア74が詳細なアノテーション情報を入力する。アノテーション情報はアノテーションサーバ73に入力され、アノテーシ

ヨンデータベース72に記録される。アノテーション情報がアノテーションデータベース72に存在する場合、その情報を加味した後処理を行うことができる。アノテーション情報がない場合には実施の形態2と同様な処理が行われる。

【0086】

図19はアノテーション用後処理モジュール71の一例を示したブロック図である。アノテーション用後処理モジュール71には差分部分マーキングモジュール75、グループ分割およびグループ選択モジュール76、グループ並び替えモジュール77を有する。

【0087】

ここでは、ページ上のビジュアルなブロックがアノテーションによって指定されているとする。まず、目的ページ6のDOMに対して、差分出力に含まれているノードにマーキングを行う(75)。次にアノテーション情報に基づいて、ページを分割し、「差分を含まないグループ」を削除する(76)。残ったグループに対しては、コンテンツをすべて復活させる(77)。これにより、グループ単位での差分を得ることができ、ビジュアルなブロックを加味した出力を得ることができる。

【0088】

(実施の形態4)

図20は、本発明の実施の形態4のシステムを示した構成図である。本実施の形態のシステムは、音声入力に対し音声でブラウジングする形態を例示する。本実施の形態のプロキシサーバ80における差分処理4は実施の形態2と同様である。また、実施の形態3のアノテーション情報(アノテーション用後処理71)との組み合わせを適用しても良い。

【0089】

本実施の形態においては、ユーザ端末はボイスXMLブラウザ81、電話82、インターネット電話83、簡易音声ブラウザ84である。また、電話82およびインターネット電話83をユーザ端末とする場合には音声認識ブラウジングサーバ85を、ボタン操作による電話82および簡易音声ブラウザ84をユーザ端末とする場合にはボタン操作音声ブラウジングサーバ86を必要とする。さらに

、プロキシサーバ80にはDOMからボイスXMLへの変換モジュール87、DOMからHTMLへの変換モジュール88を有する。

【0090】

ボイスXMLブラウザ81はクライアントサイドで稼動する音声応答を実現するソフトウェア機能である。電話82では音声入力またはボタン操作入力に対する音声応答を得、インターネット電話83では音声入力に対する音声応答を得る。簡易音声ブラウザ84は、キー操作入力に対し音声出力を得る。

【0091】

音声認識ブラウジングサーバ85は、音声合成エンジンと音声認識エンジンとを有する。電話82またはインターネット電話83から入力された音声を音声認識エンジンで認識し、入力内容を解析する。入力内容はコマンドまたはデータとしてプロキシサーバ80に送付される。プロキシサーバ80から受け取ったボイスXMLデータを音声合成エンジンによって音声に変換し、ユーザサイドの端末（電話またはインターネット電話）で再生する。

【0092】

ボタン操作音声ブラウジングサーバ86は、電話82からのボタン操作入力をDTMF信号処理部で処理しコマンドまたはデータに変換する。また、簡易音声ブラウザ84からのキー操作をキー操作処理部で解析しコマンドまたはデータに変換する。コマンドまたはデータはプロキシサーバ80に送付される。プロキシサーバ80からの応答はHTMLファイルとして受け取り、音声合成エンジンがこれを音声に変換してユーザ端末である電話82または簡易音声ブラウザ84で再生する。

【0093】

図21は、代表的な音声ブラウザの構成を示した図である。HTTPリクエストに対する応答として受け取ったHTML文書（ファイル）をパージング・解析モジュール90で解析し、読み上げ情報91に変換する。読み上げ情報91は読み上げ制御モジュール92で制御されて、音声合成エンジン93によって音声信号に合成される。合成された音声はスピーカ94から出力される。なお、読み上げ制御モジュール92はキーボード等の入力デバイス95からの入力を受け付け

て読み上げ情報を制御できる。

【0094】

図22は、ボイスXML用の後処理モジュール89およびDOMからボイスXMLへの変換モジュール87の一例を示したブロック図である。リストタイトル修復モジュール64、表頭・表側修復モジュール65、フォーム移動モジュール66については実施の形態2と同様である。但し本実施の形態では、各モジュールにおいて修復・移動を行うと同時にリストタイトル96、テーブルタイトル97、フォームタイトル98をリストアップしタイトルリスト99を生成する。このタイトルはそれぞれのタイトルの内容を「かたまり」としてとらえ、このかたまりの指標であると考えることができる。たとえばリストタイトルはリンクリストを1つのかたまりとして、フォームであれば1つのフォームを1つのかたまりとして考えることができる。これら「かたまり」を認識することによりHTMLファイルからの音声応答を生成できる。なお、この際に実施の形態3で説明したアノテーション情報により音声合成精度を高めることができる。

【0095】

得られたタイトルリストに対してキーワード解析100を施し、音声認識用のボキャブラリを生成する(102)。一方DOMツリーとして得られている差分ページからリンク(アンカータグ)をリストアップし(101)、キーワード解析100を施して音声認識用のボキャブラリを生成する(102)。得られたボキャブラリから応答文および文法を生成する(103)。

【0096】

本実施の形態のシステムおよび方法によれば、音声入力あるいは簡単なキー操作入力(ボタン入力)によって、ウェブページの閲覧を音声出力により得ることができる。視覚障害者がウェブコンテンツにアクセスする場合にバリアフリーを実現する有効な手段を提供できる。なお、コンテンツの内容が簡略化されるので音声応答による読み上げがスムーズに進むのは勿論である。また、コンピュータの操作に不慣れなユーザに対しても同様に簡便なウェブコンテンツへのアクセスを可能にする手法を提供できる。

【0097】

以上、本発明者によってなされた発明を発明の実施の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることは言うまでもない。

【0098】

【発明の効果】

本願で開示される発明のうち、代表的なものによって得られる効果は、以下の通りである。すなわち、小画面デバイスや音声ブラウザを用いたウェブページの表示あるいは出力の際に、必要な情報に迅速にアクセスするためのウェブページ簡略化の手法を提供できる。また、ウェブページの簡略化を同一URLの過去のページが存在しない場合においても実行できる。また、ウェブページの簡略化をオンザフライで実行できる。また、ウェブページの簡略化の際に重要な情報を欠落することなく、不必要的情報を精度良く簡略化する手法を提供できる。

【図面の簡単な説明】

【図1】

本発明の実施の形態1のシステムを示した構成図である。

【図2】

差分処理の詳細を示したブロック図である。

【図3】

隣接URL列挙モジュールの詳細を示したブロック図である。

【図4】

URLキャッシュモジュールの一例を示した構成図である。

【図5】

URL優先順位演算モジュールの一例を示したブロック図である。

【図6】

更新前目的ページ/隣接ページ取得モジュールの一例を示した構成図である。

【図7】

フェッチモジュールの一例を示す構成図である。

【図8】

差分演算モジュールの一例を示したブロック図である。

【図9】

ノード位置による重要度の相違を示す図である。

【図10】

通常のブラウザによりニュースページを表示した表示図である。

【図11】

本実施の形態のシステムを用いて図10のページを表示した表示図である。

【図12】

検索画面の一例を通常のブラウザにより表示した例を示す表示図である。

【図13】

本実施の形態のシステムを用いて図12のページを表示した表示図である。

【図14】

本発明の実施の形態2のシステムを示した構成図である。

【図15】

後処理モジュールの一例を示したブロック図である。

【図16】

リストタイトル修復モジュールによる修復の例を示す表示図である。

【図17】

(a) はフォームの存在した位置に移動したフォームへのリンクを残した図であり、図17 (b) はページ下部に移動されたフォームを示す図である。

【図18】

本発明の実施の形態3のシステムを示した構成図である。

【図19】

アノテーション用後処理モジュールの一例を示したブロック図である。

【図20】

本発明の実施の形態4のシステムを示した構成図である。

【図21】

代表的な音声ブラウザの構成を示した図である。

【図22】

ボイスXML用の後処理モジュールおよびDOMからボイスXMLへの変換モ

ジユールの一例を示したブロック図である。

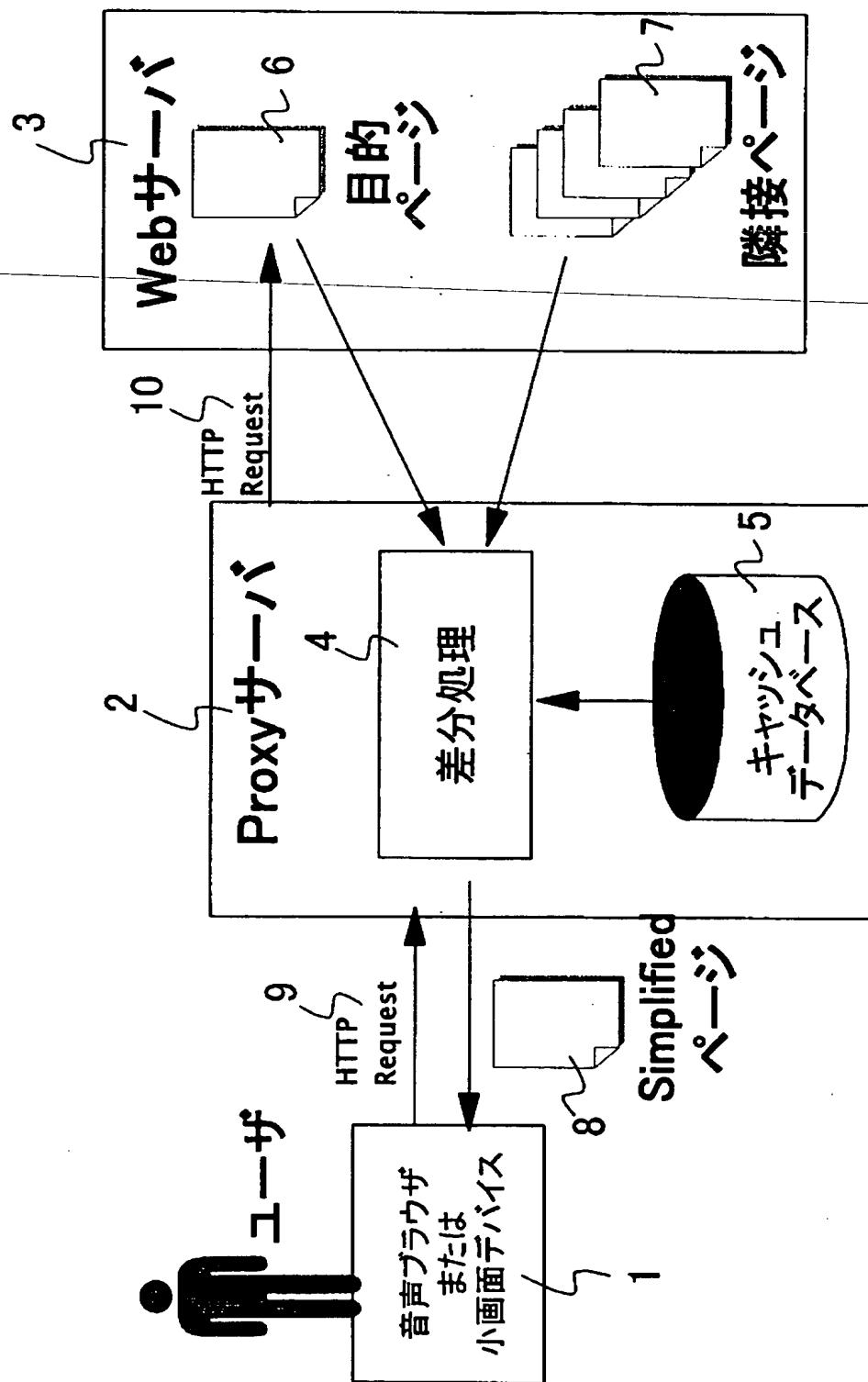
【符号の説明】

1 …ユーザ端末、 2 …プロキシサーバ、 3 …ウェブサーバ、 4 …差分処理、 5 …キャッシュデータベース、 6 …目的ページ、 7 …隣接ページ、 8 …簡略化ページ、 9 …HTTPリクエスト、 10 …HTTPリクエスト、 11 …隣接URL列挙モジュール、 12 …ディレクトリ列挙モジュール、 13 …URLキャッシュモジュール、 14 …更新前目的ページ／隣接ページ取得モジュール、 15 …フェッチモジュール、 16 …URL優先順位演算モジュール、 17 …差分演算モジュール、 18 …最小差分選択モジュール、 19 …リンク列挙モジュール、 20 …URLリスト、 21 …ディレクトリ共通URL選択モジュール、 22 …親ディレクトリ共通URL選択モジュール、 23 …マージモジュール、 24 …隣接URLリスト、 25 …URLキャッシュデータベース、 26 …検索手段、 27 …URLリスト、 28 …同一ページ除外モジュール、 29 …URLエディットディスタンス算出モジュール、 30 …URL関連性算出モジュール、 31 …ソートモジュール、 32 …リダイレクションURLテーブル、 33 …URL共起回数テーブル、 34 …URL相互参照テーブル、 35 …優先度順URLリスト、 36 …検索キー作成モジュール、 37 …検索モジュール、 38 …リスト、 39 …HTMLパーザ、 40 …ウェブページ、 41 …更新前目的ページ／隣接ページリスト、 42 …ダウンロードモジュール、 43 …HTMLパーザ、 45 …HTMLファイル、 46 …DOMツリー、 47 …線形化モジュール、 48, 49 …ノードリスト、 50 …DPマッチングモジュール、 51 …共通ノードリスト、 52 …重要度算出モジュール、 53 …共通ノード削除モジュール、 54 …クリーンアップモジュール、 55 …差分DOMツリー、 60 …プロキシサーバ、 61 …後処理モジュール、 62 …変換モジュール、 63 …ユーザプロファイル、 64 …リストタイトル修復モジュール、 65 …表頭・表側修復モジュール、 66 …フォーム移動モジュール、 70 …プロキシサーバ、 71 …アノテーション用後処理モジュール、 72 …アノテーションデータベース、 73 …アノテーションサーバ、 74 …ボランティア、 75 …差分部分マーキングモジュール、 76 …グループ選択モジュール、 77 …グループ並び替えモジュール、 80 …プロキシサーバ、 81 …ボイスXMLブラウザ、

82…電話、83…インターネット電話、84…簡易音声ブラウザ、85…音声認識ブラウジングサーバ、86…ボタン操作音声ブラウジングサーバ、87…変換モジュール、88…変換モジュール、89…後処理モジュール、90…ページング・解析モジュール、91…読み上げ情報、92…読み上げ制御モジュール、93…音声合成エンジン、94…スピーカ、95…入力デバイス、96…リストタイトル、97…テーブルタイトル、98…フォームタイトル、99…リストアップしタイトルリスト、100…キーワード解析。

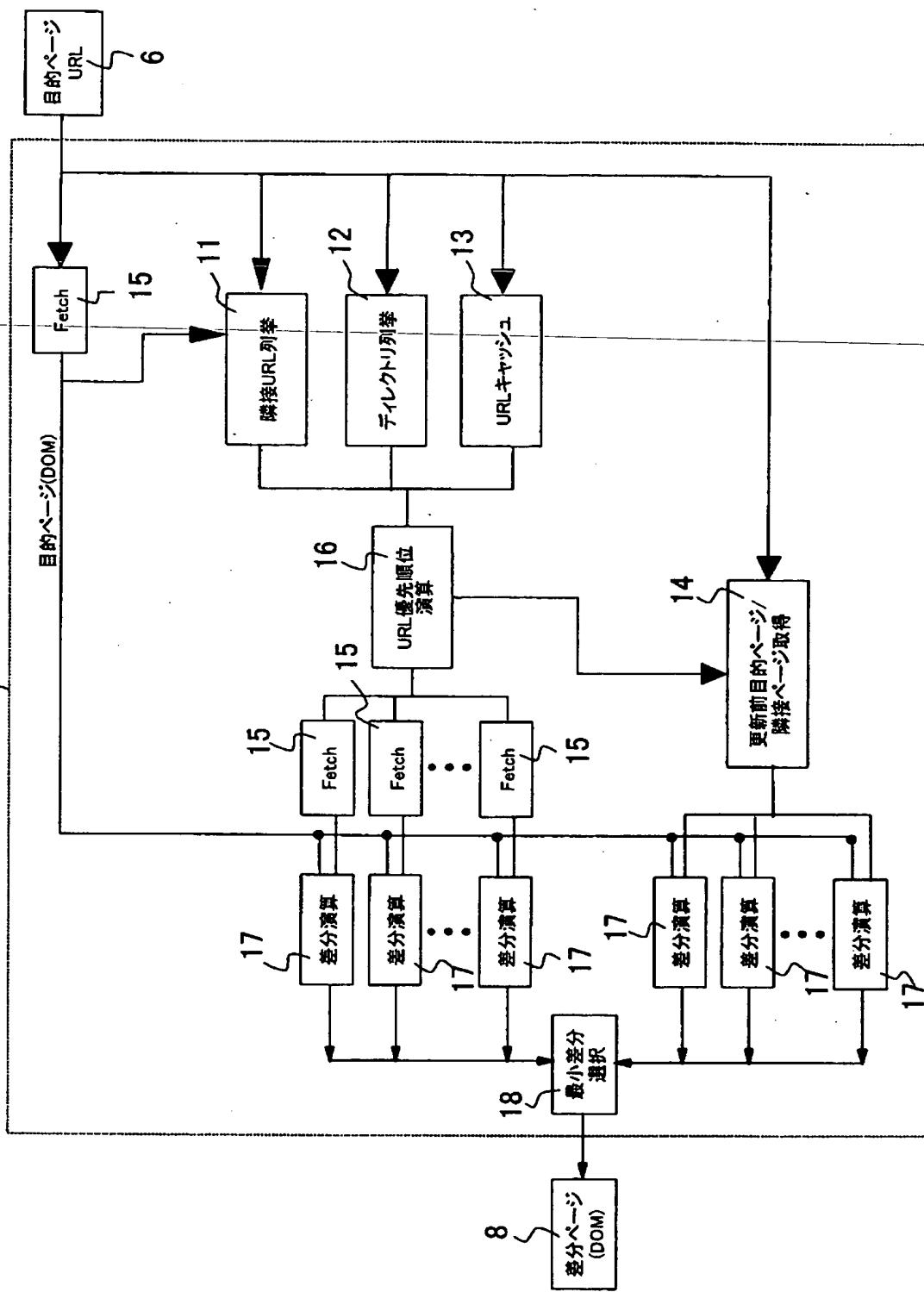
【書類名】 図面

【図1】

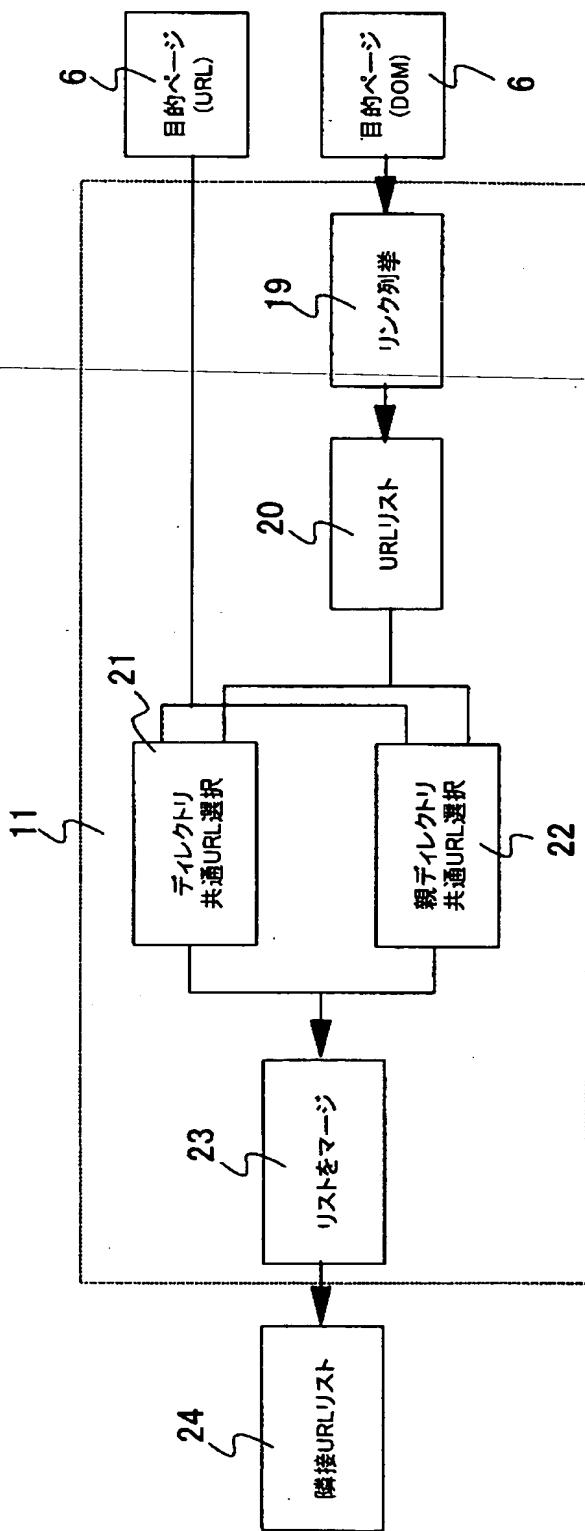


【図2】

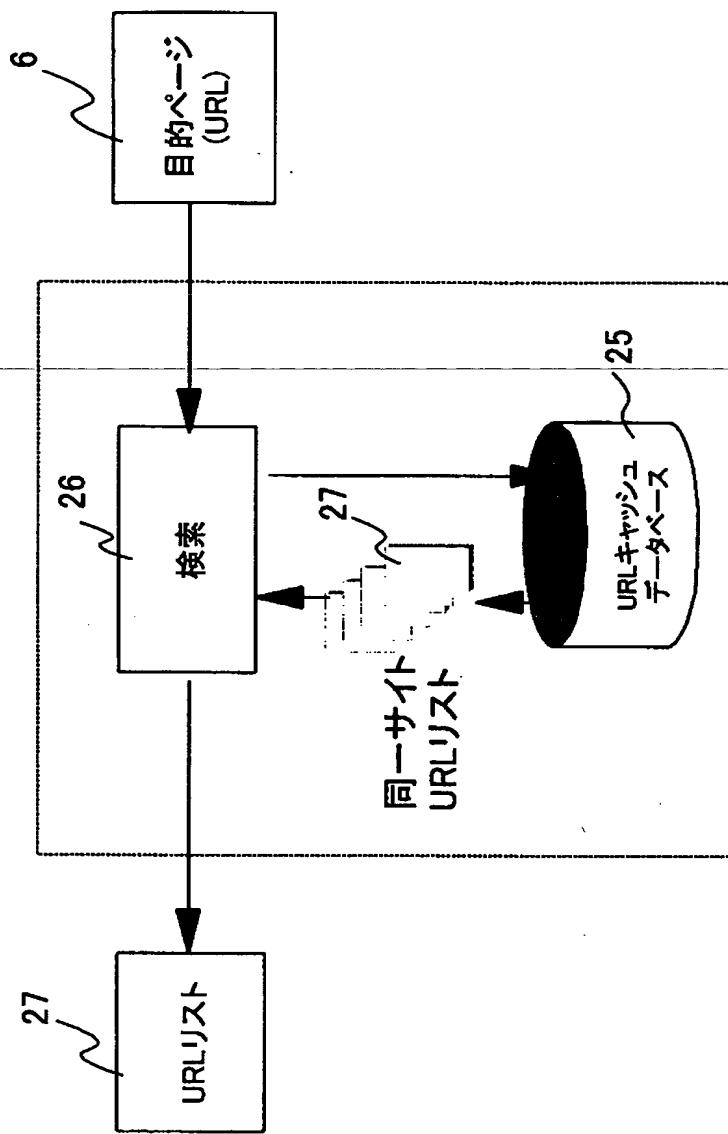
4



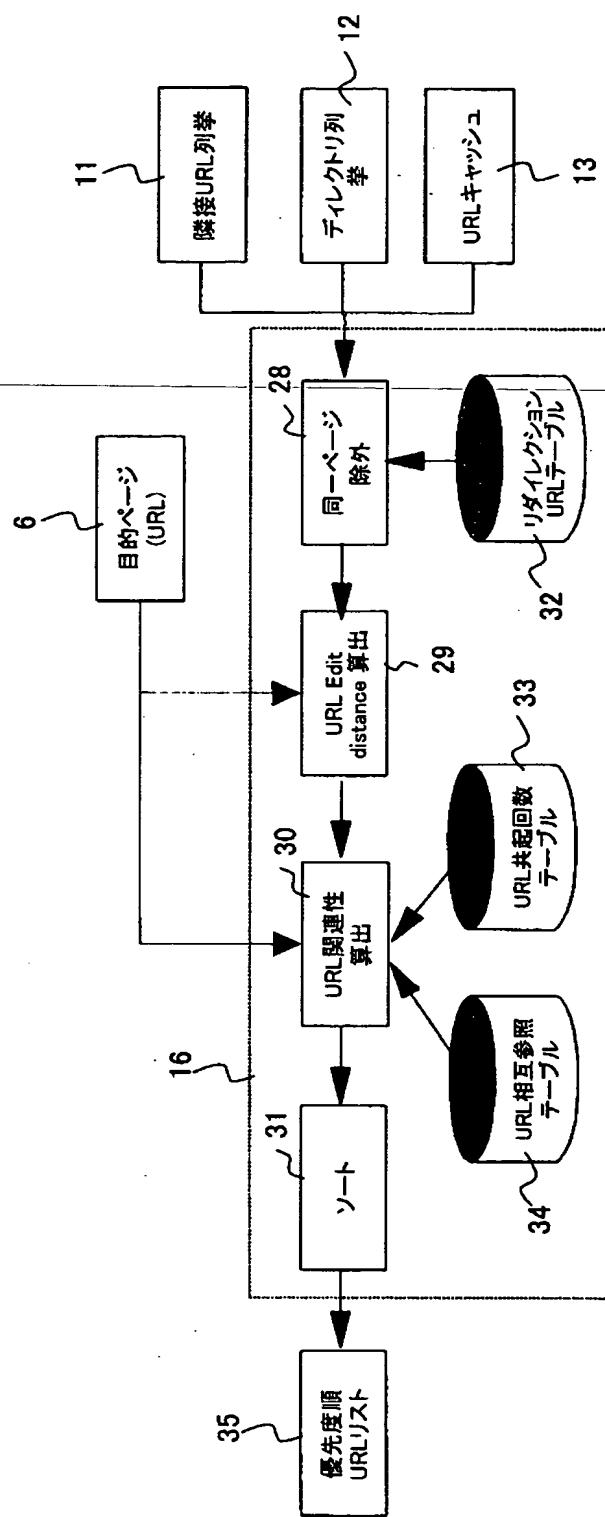
【図3】



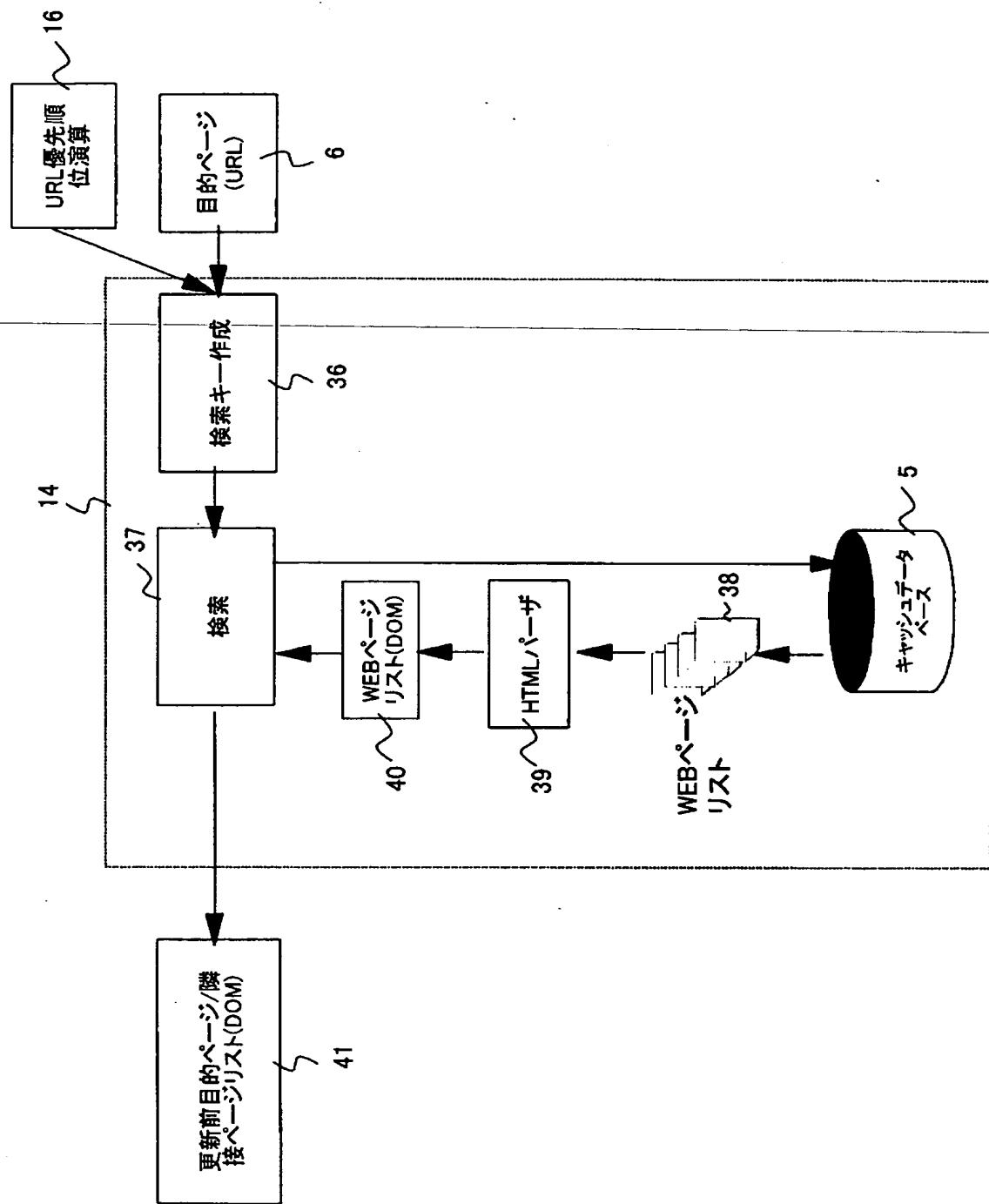
【図4】



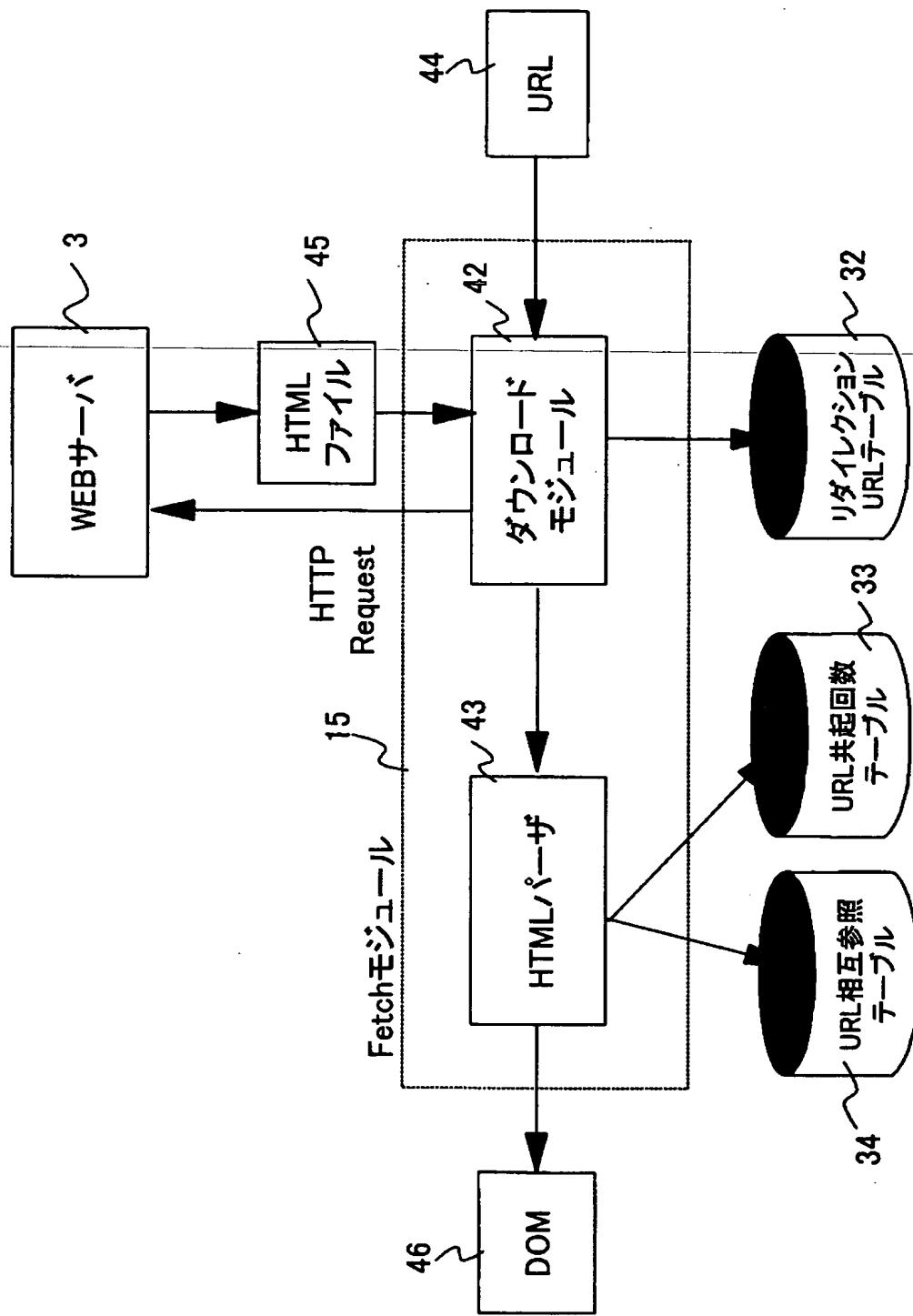
【図5】



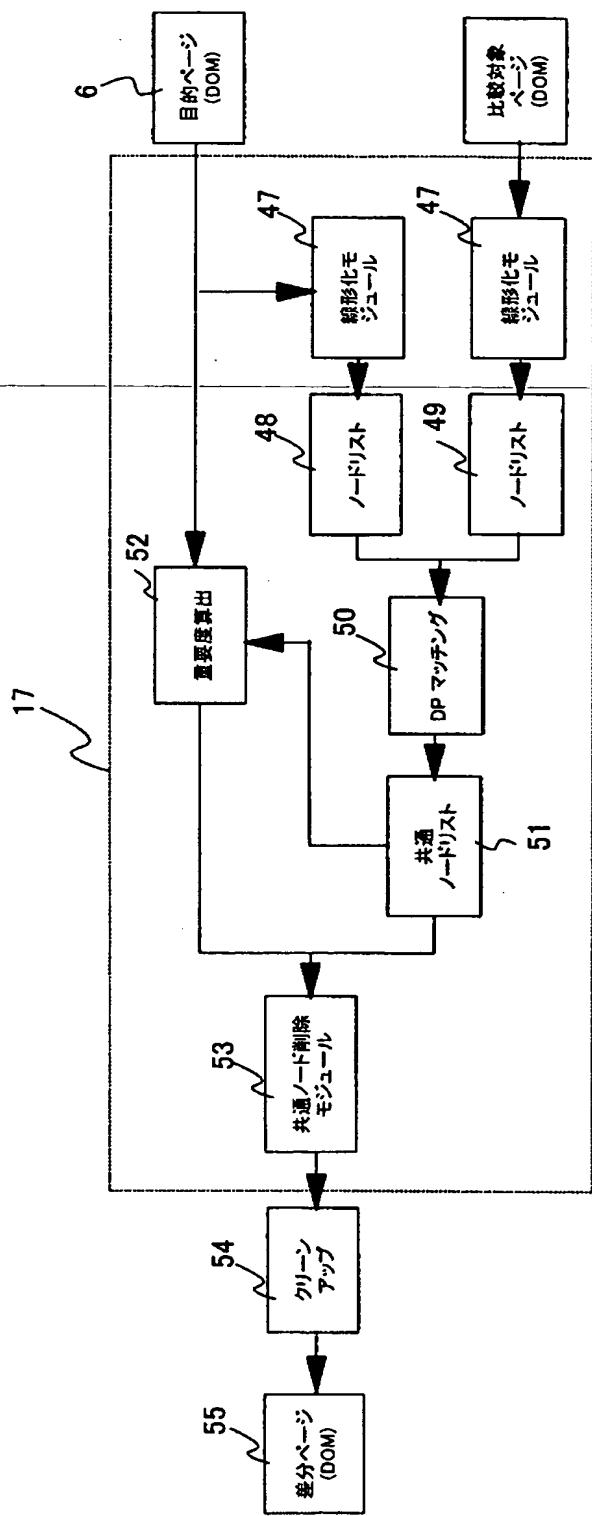
【図6】



【図7】

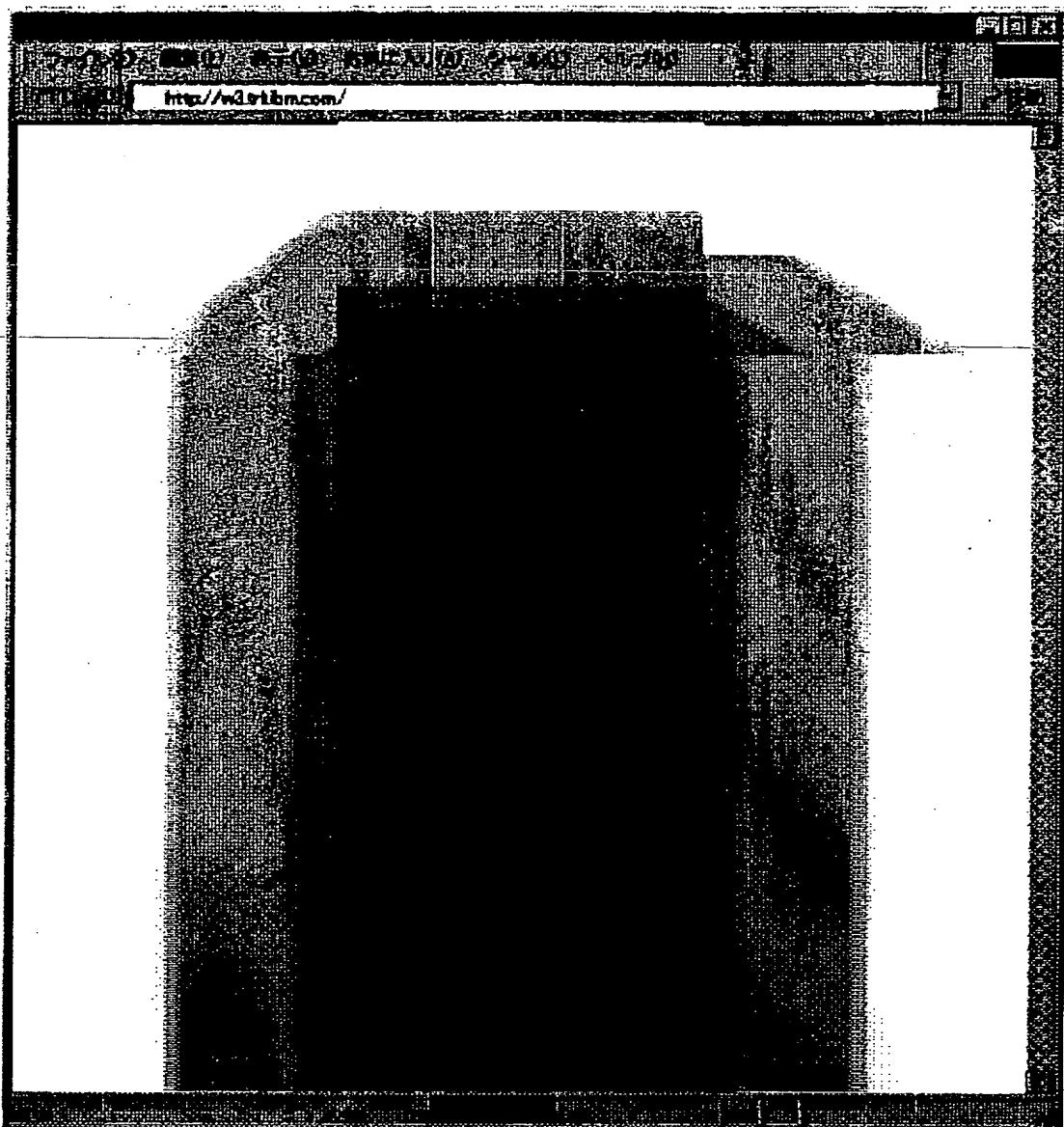


【図8】

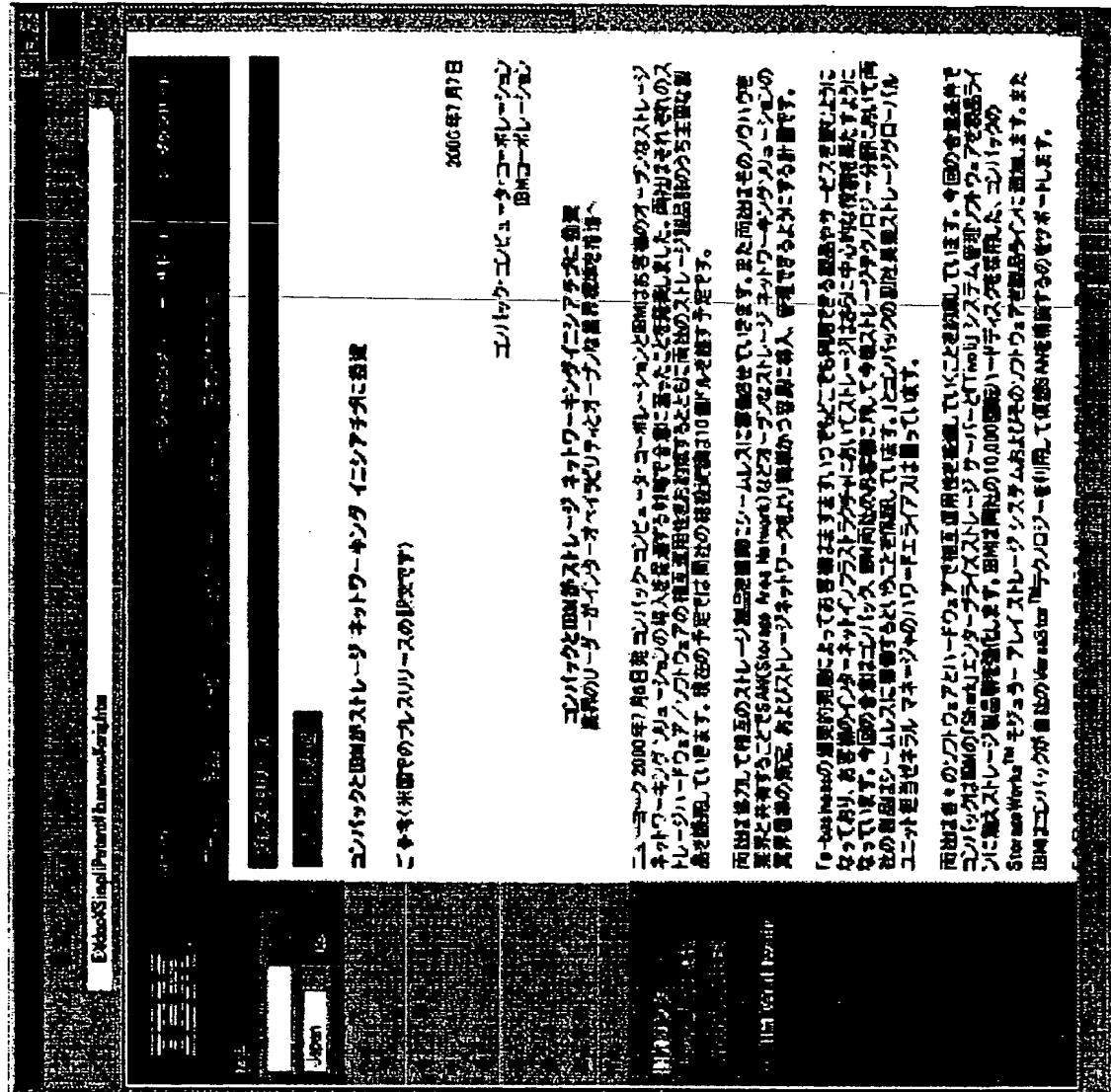


特2000-227996

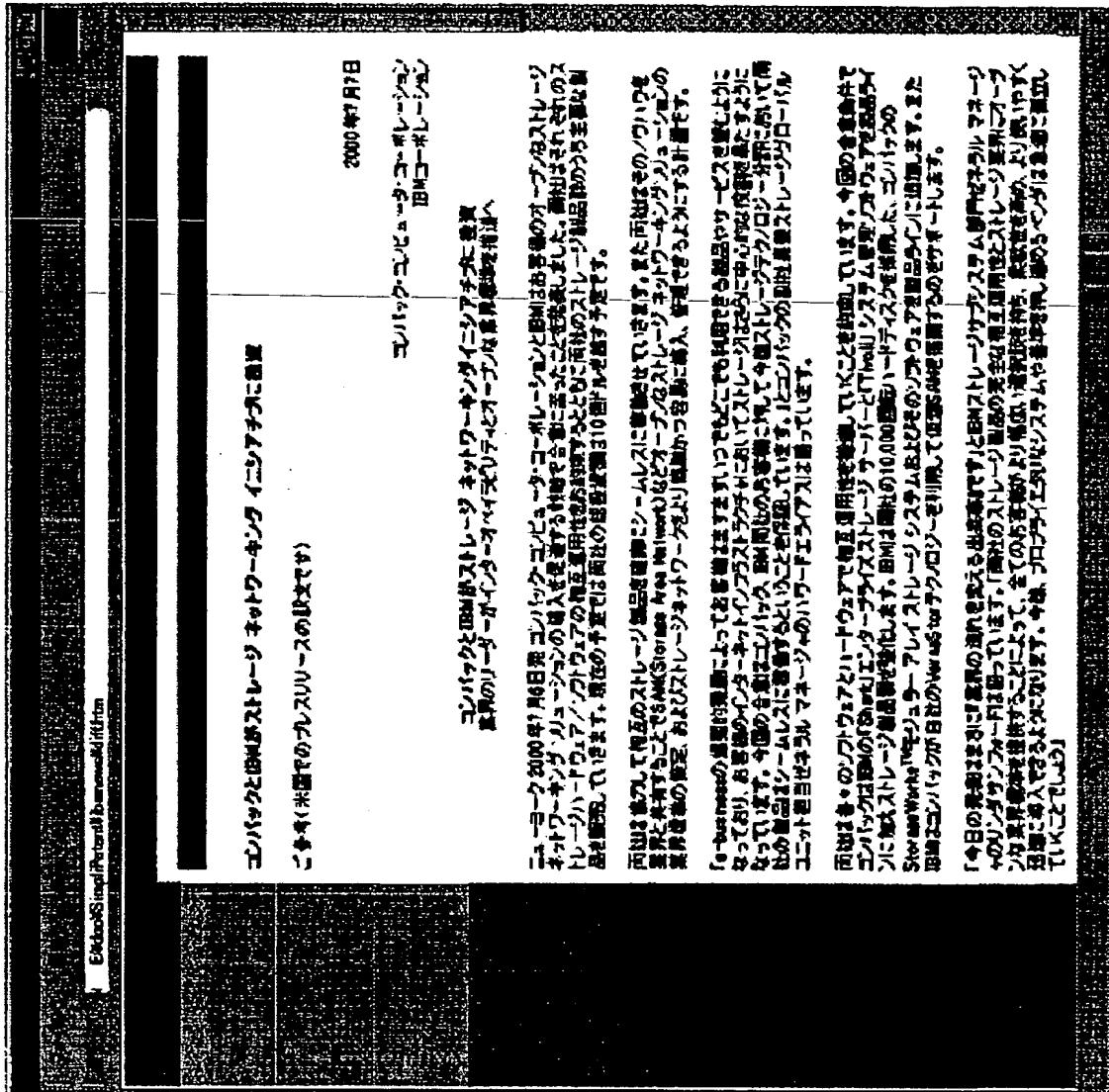
【図9】



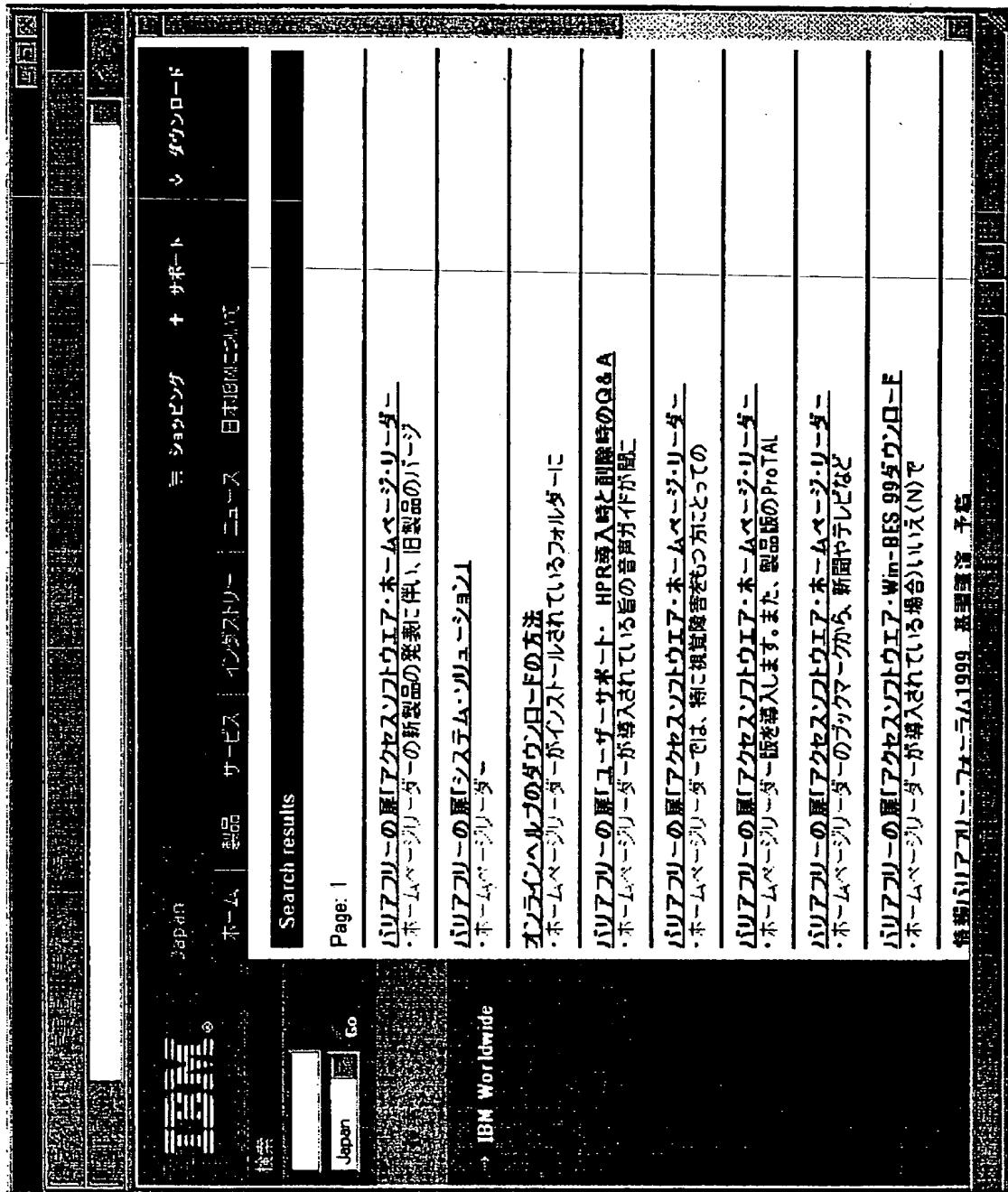
【図10】



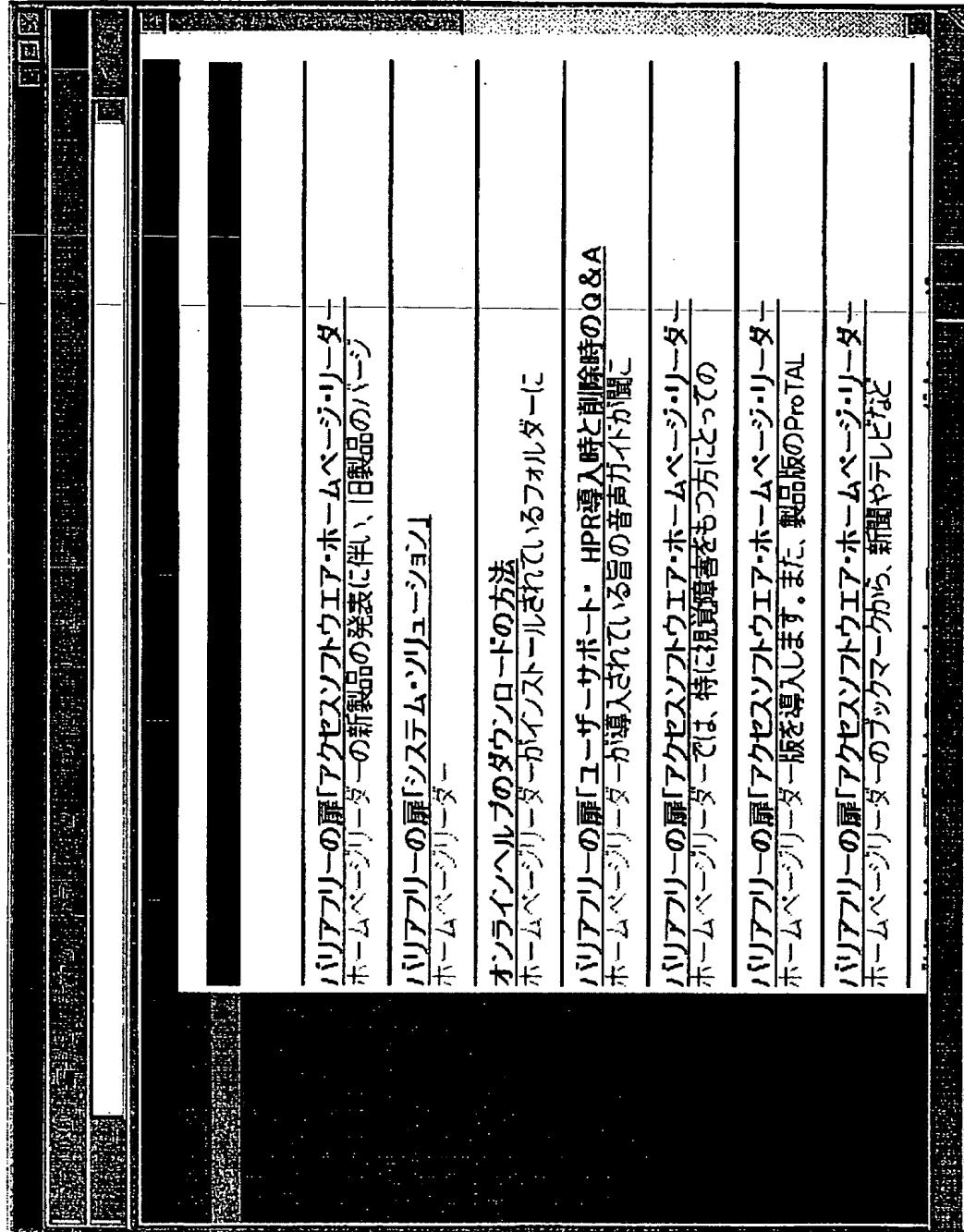
【圖 1 1】



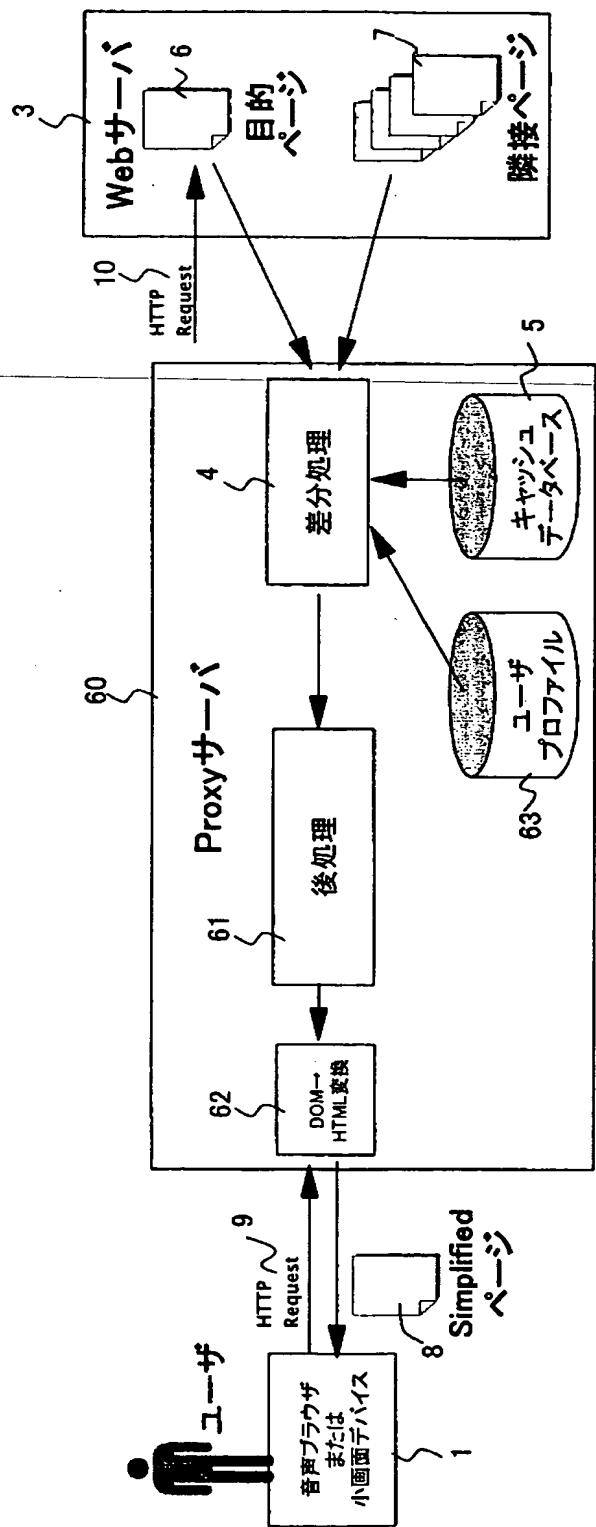
【図12】



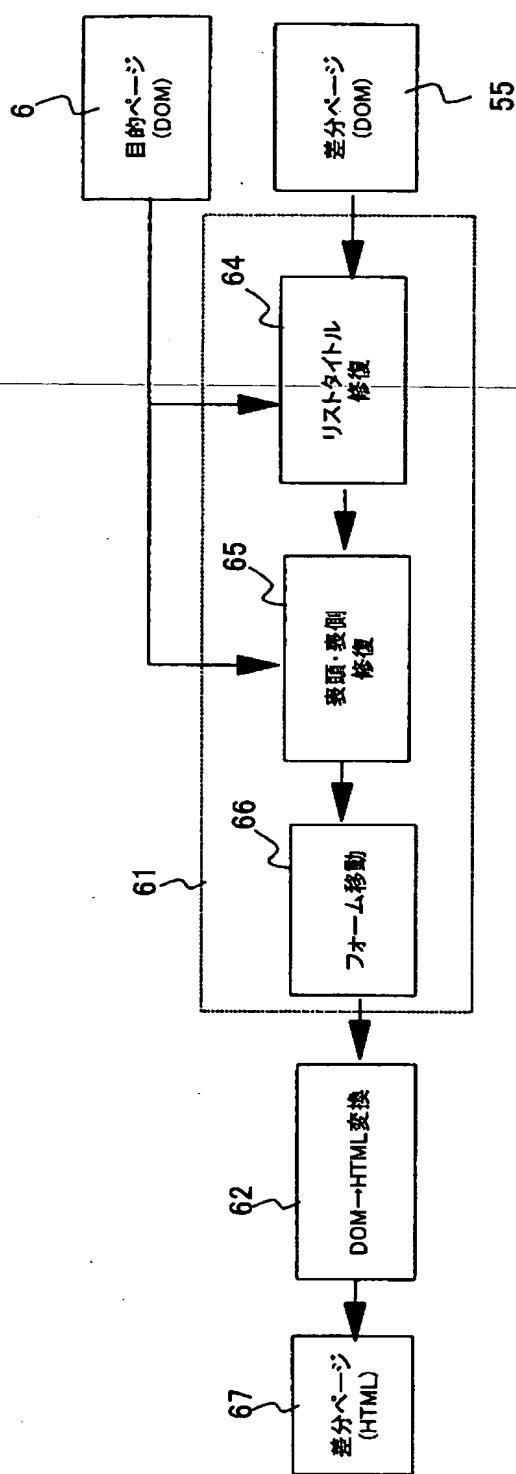
【図13】



【図14】



【図15】



【図16】

(a)

同ジャンルの他のニュース

- 半導体メーカー、デジタル需要増見込み設備投資を積極化(14:15)
- 太陽電池の生産、本家の米国拠点で日本が世界一(12:50)
- 企業会計ルールの統一、欧米主導に日本危機感(11:12)
- 祝金や公共料金の支払い、電話・ネットでも可能に(23:37)
- TTNetが学校を対象に割安サービス開始 5月から(20:24)
- ムルノーが三星自動車を買収(13:33)

(b)

- 半導体メーカー、デジタル需要増見込み設備投資を積極化(14:15)
- 企業会計ルールの統一、欧米主導に日本危機感(11:12)

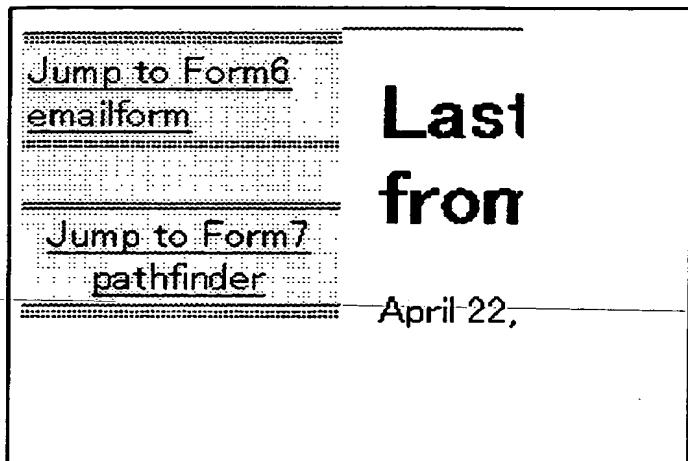
(c)

同ジャンルの他のニュース

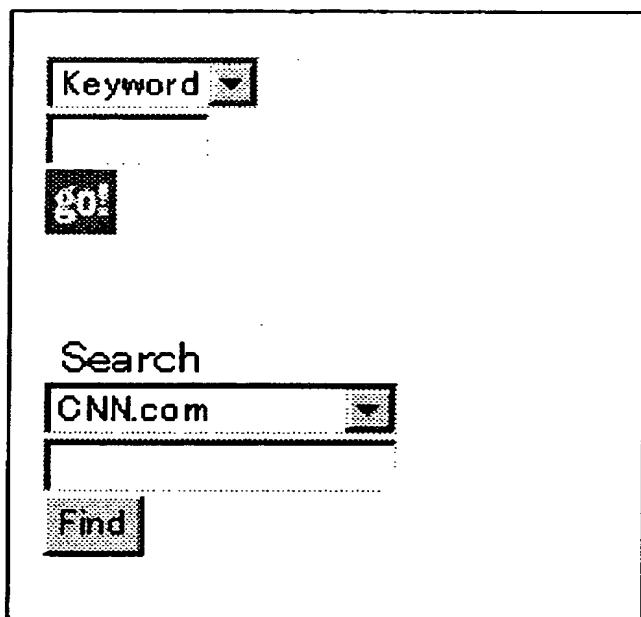
- 半導体メーカー、デジタル需要増見込み設備投資を積極化(14:15)
- 企業会計ルールの統一、欧米主導に日本危機感(11:12)

【図17】

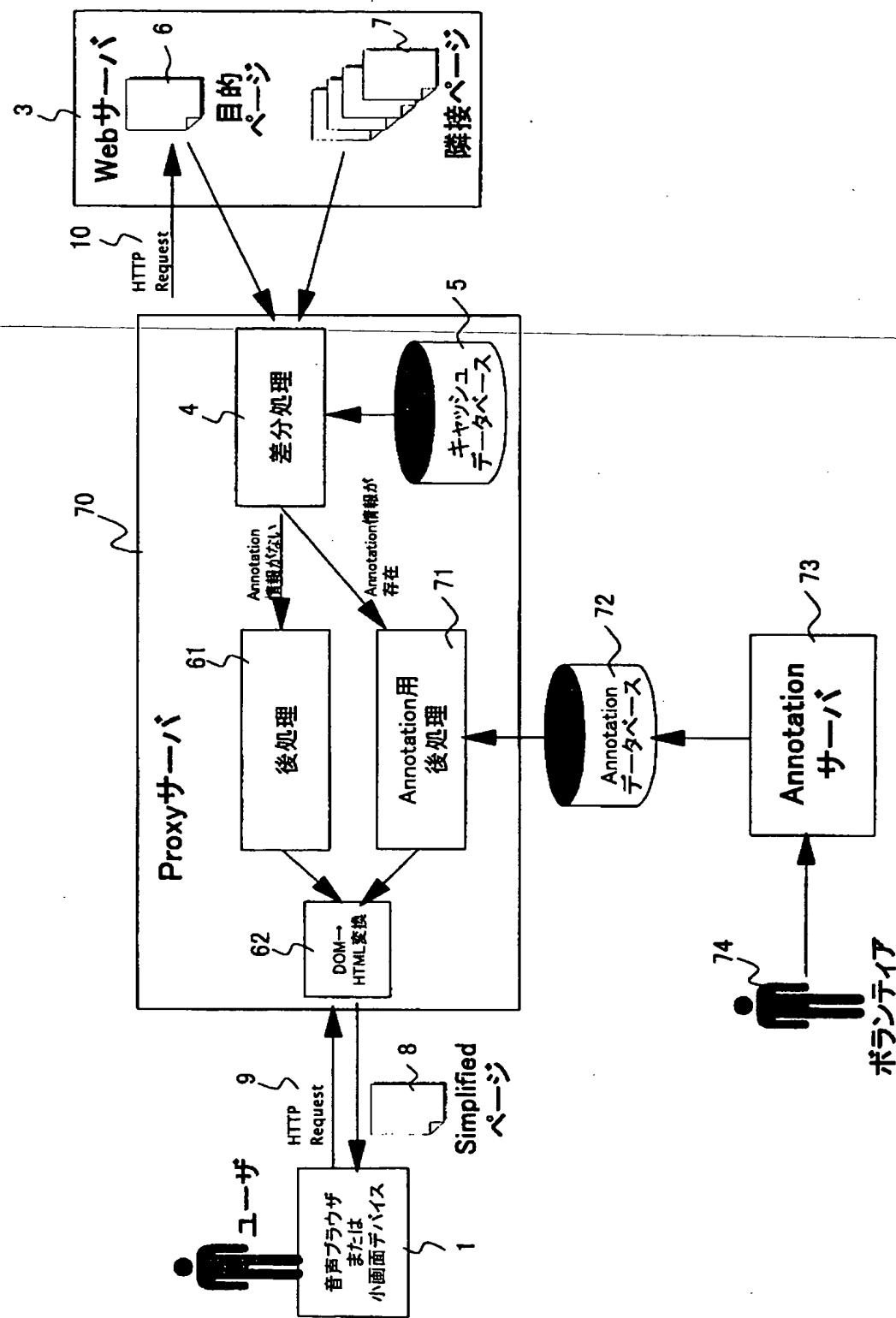
(a)



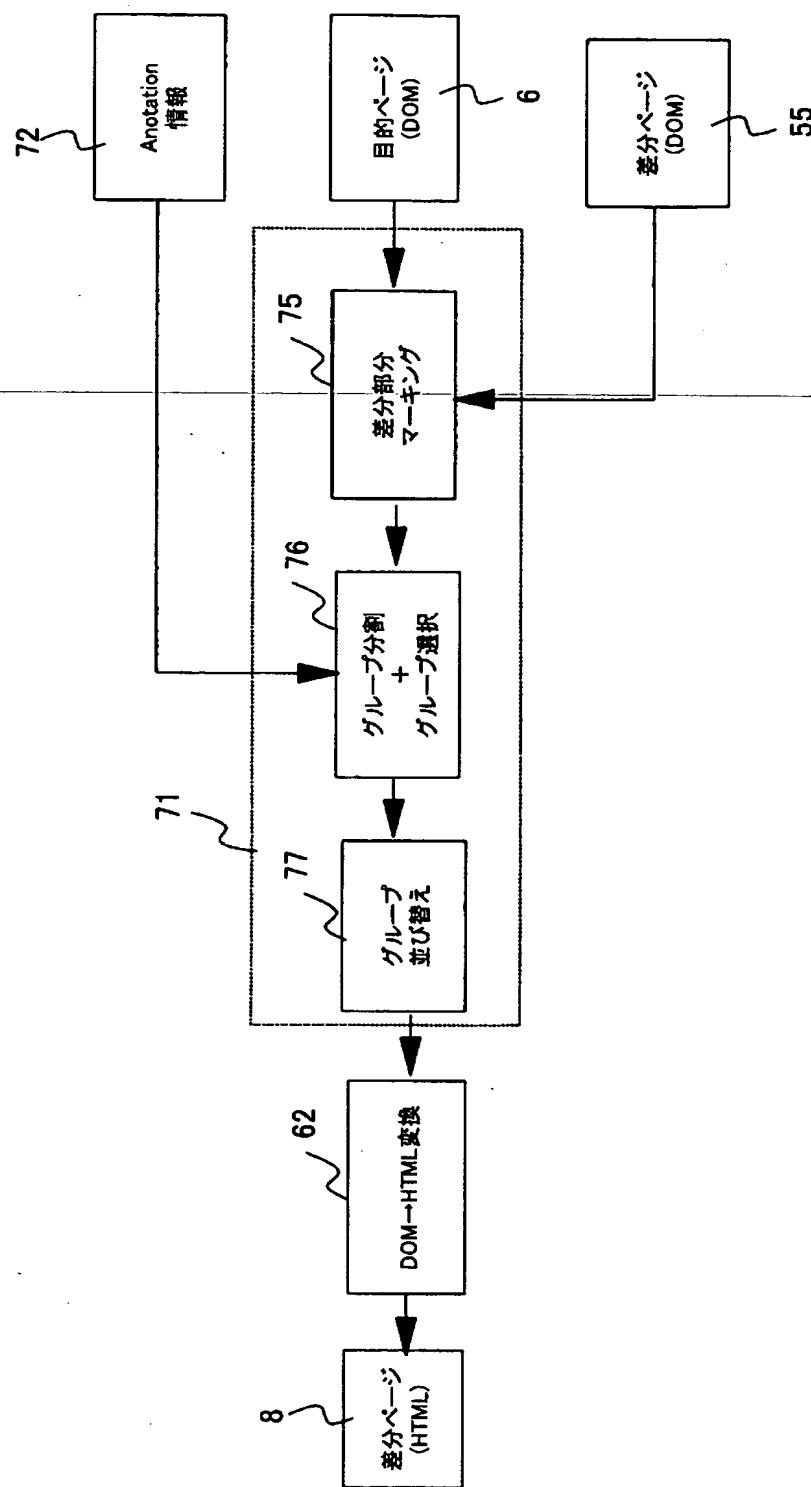
(b)



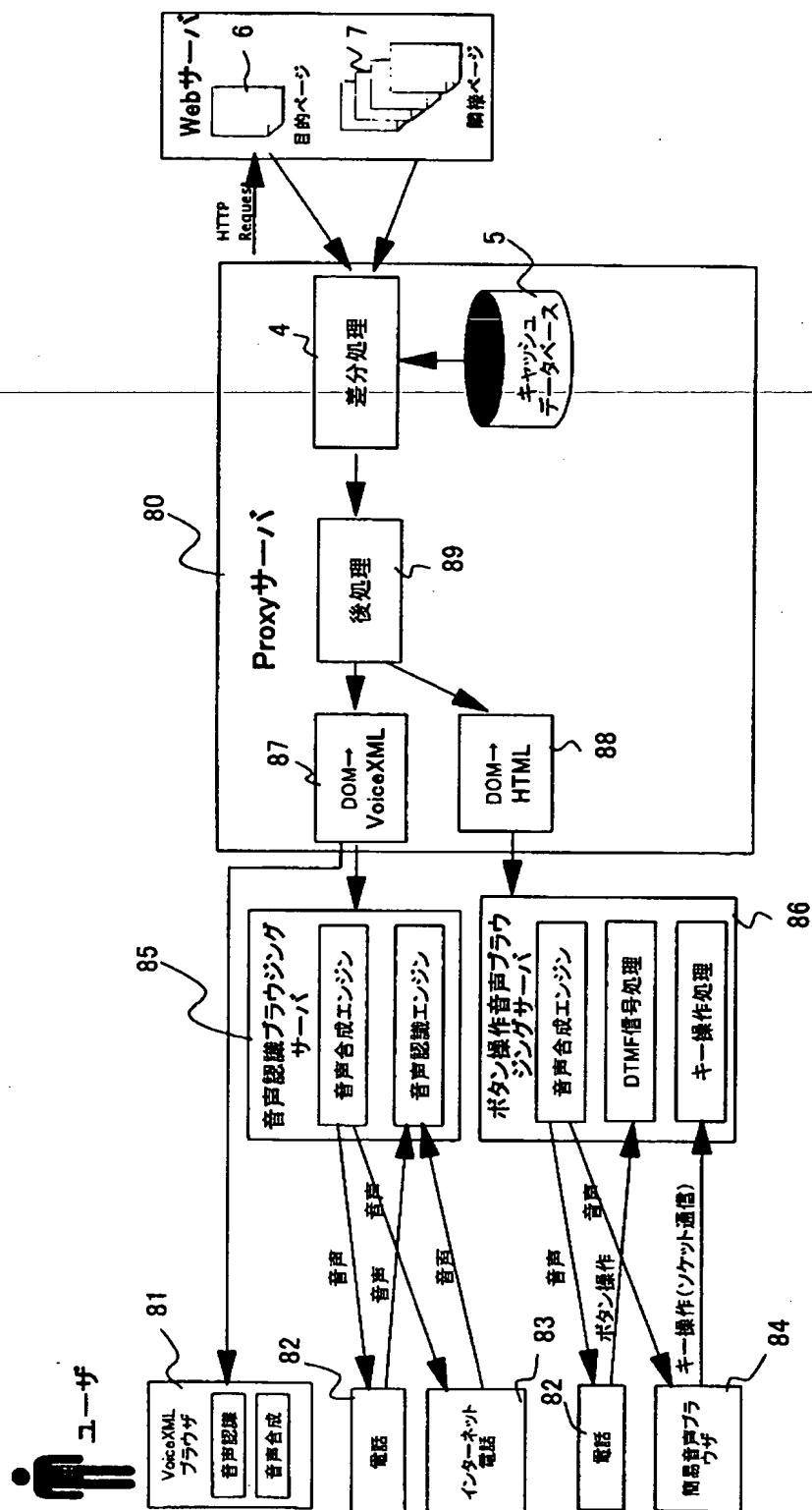
【図18】



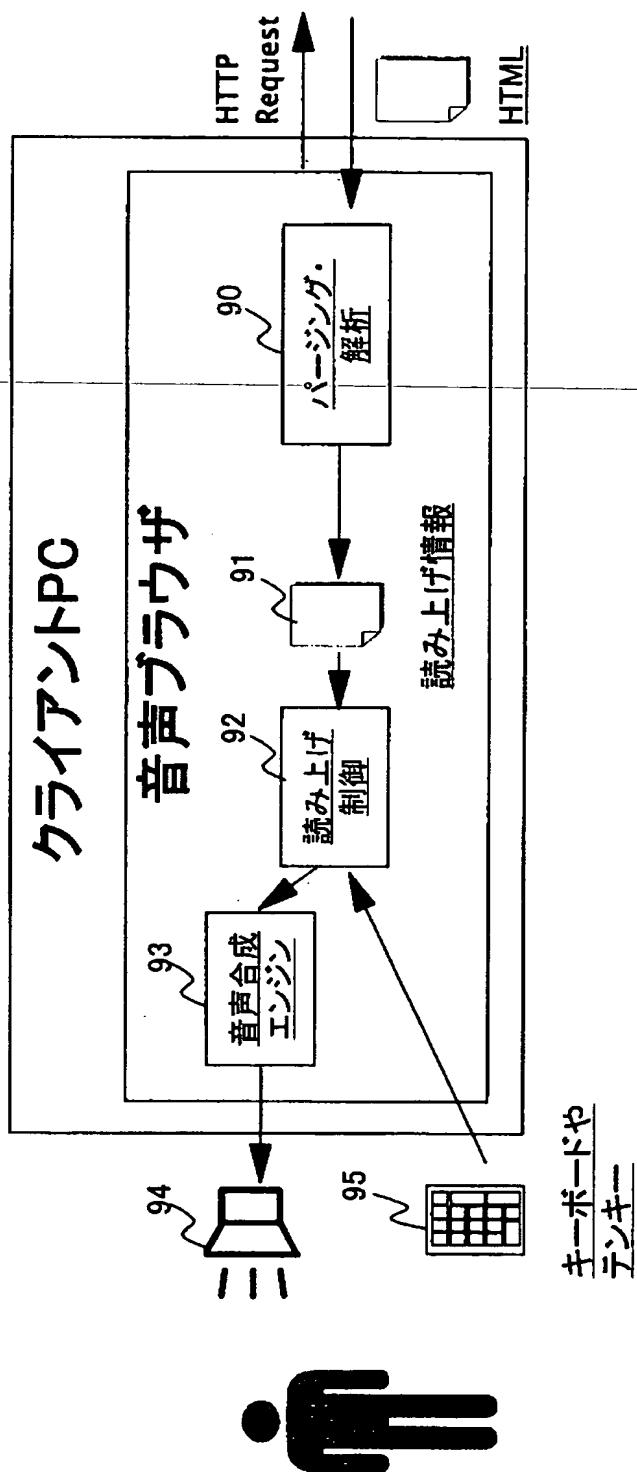
【図19】



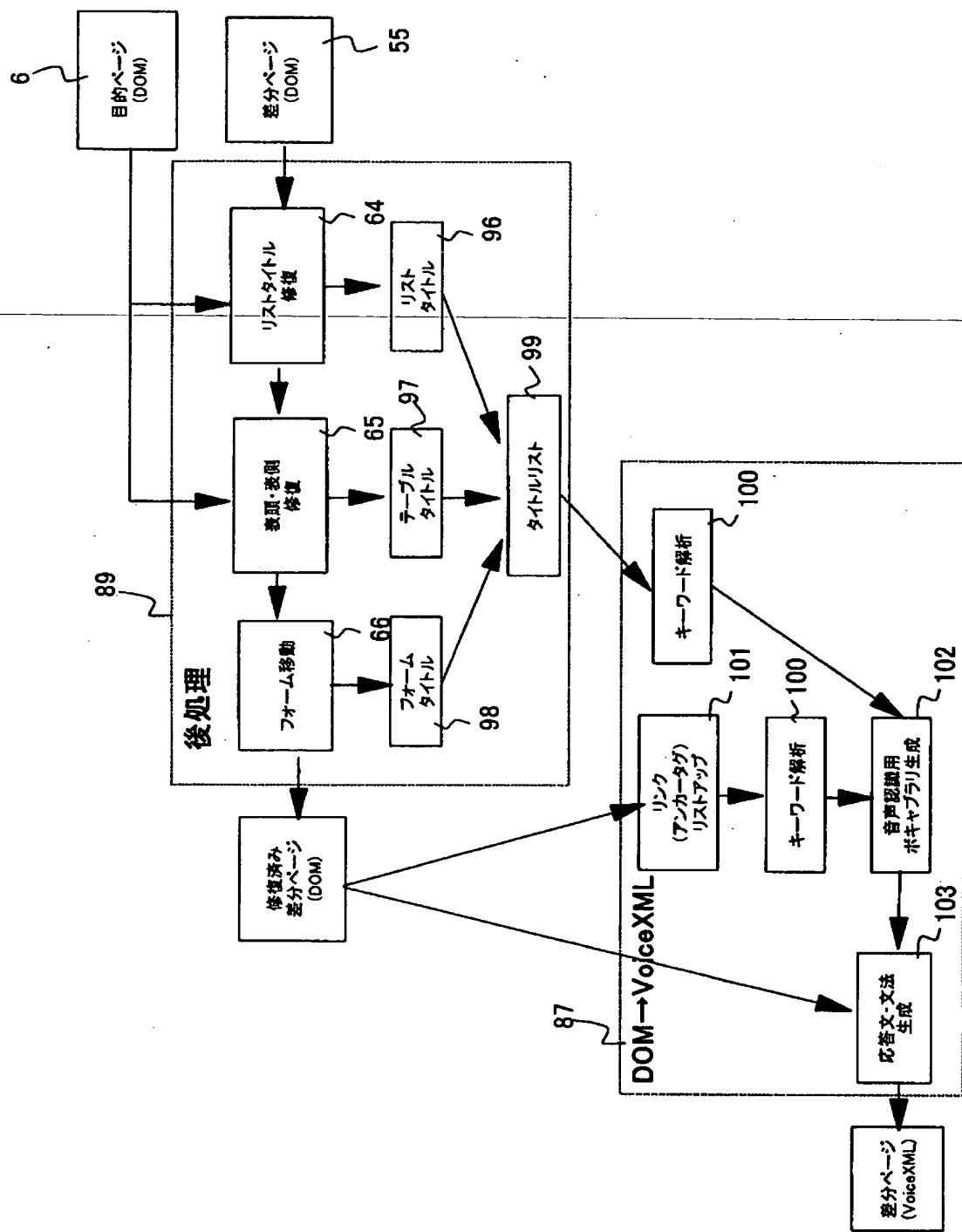
【図20】



【図21】



【図22】



【書類名】 要約書

【要約】

【課題】 小画面デバイスや音声ブラウザを用いたウェブページの表示あるいは出力の際に、必要な情報に迅速にアクセスするためのウェブページ簡略化の手法を提供する。

【解決手段】 ページ簡略化で用いる差分処理4の計算では、簡略化対象の目的ページ6と比較する比較ページとして、同一URLの過去ページのみならず、隣接ページ7を用いる。比較により更新された情報のみを取り出し、サイト内の各ページで共通に使用されているテンプレート情報を取り除く。これにより簡略化対象ページの主なコンテンツのみを取り出した簡略化ページ8を得る。

【選択図】 図1

認定・付加情報

特許出願の番号 特願2000-227996
 受付番号 50000953910
 書類名 特許願
 担当官 濱谷 よし子 1614
 作成日 平成12年 9月 6日

<認定情報・付加情報>

【特許出願人】

【識別番号】 390009531
 【住所又は居所】 アメリカ合衆国 10504、ニューヨーク州 アーモンク (番地なし)
 【氏名又は名称】 インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】 100086243
 【住所又は居所】 神奈川県大和市下鶴間 1623 番地 14 日本アイ・ビー・エム株式会社 大和事業所内

【氏名又は名称】 坂口 博

【復代理人】

【識別番号】 100112520
 【住所又は居所】 神奈川県大和市中央林間 3 丁目 4 番 4 号 サクライビル 4 階 間山・林合同技術特許事務所
 【氏名又は名称】 林 茂則

【選任した代理人】

【識別番号】 100091568
 【住所又は居所】 神奈川県大和市下鶴間 1623 番地 14 日本アイ・ビー・エム株式会社 大和事業所内
 【氏名又は名称】 市位 嘉宏

【選任した代理人】

【識別番号】 100106699
 【住所又は居所】 神奈川県大和市下鶴間 1623 番 14 日本アイ・ビー・エム株式会社 大和事業所内
 【氏名又は名称】 渡部 弘道

【選任した復代理人】

【識別番号】 100110607

次頁有

認定・付加情報 (続巻)

【住所又は居所】 神奈川県大和市中央林間3丁目4番4号 サクラ
イビル4階 間山・林合同技術特許事務所
【氏名又は名称】 間山 進也
【選任した復代理人】
【識別番号】 100098121
【住所又は居所】 神奈川県大和市中央林間3丁目4番4号 サクラ
イビル4階 間山・林合同技術特許事務所
【氏名又は名称】 間山 世津子

次頁無

出願人履歴情報

識別番号 [390009531]

1. 変更年月日 2000年 5月16日

[変更理由] 名称変更

住 所 アメリカ合衆国 10504、ニューヨーク州 アーモンク (番地なし)

氏 名 インターナショナル・ビジネス・マシーンズ・コーポレーション